Computational mapping of regulatory domains of human genes

D I S S E R T A T I O N
zur Erlangung des akademischen Grades
Doctor of Philosophy (Ph.D.)

eingereicht an der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
von
M.Sc. Inga Patarčić

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen:
1. Prof. Dr. Ana Pombo
2. Prof. Dr. Xianjun Dong
3. Dr. Altuna Akalin

Tag der mündlichen Prüfung: 22. Juni 2021

Contents

Erklärung

Ich erkläre, dass diese Doktorarbeit mit dem Titel "**Computational Mapping of regulatory domains of human genes**" allein von mir verfasst worden ist. Sie wurde bei keiner früheren Bewerbung um einen akademischen Grad, weder ganz noch teilweise, abgegeben.

Declaration

I declare that my PhD thesis "***Computational mapping of regulatory domains of human genes"*** has been composed solely by myself. It has not been submitted, in whole or in part, in any previous application for a degree.

Inga Patarčić

In Berlin, September 2020

The work discussed in this thesis was performed between the years 2015-2020 in the laboratory of. Dr. Altuna Akalin (Bioinformatics Platform) at the Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association in Berlin, Germany.

Abstract

Human genome contains millions of regulatory elements - enhancers - that quantitatively regulate gene expression. Multiple experimental and computational approaches have been developed to associate enhancers with their gene targets. Despite the tremendous progress in understanding how enhancers tune gene expression, the field still lacks an approach that is systematic, integrative and accessible for discovering and documenting *cis*-regulatory relationships across the genome.

To address this challenge, we developed a novel computational approach that models and integrates *gene expression ~ enhancer activity* (reg2gene). reg2gene was built upon extensive data modeling and integration, and as such, it consists of three main steps: 1) data quantification, 2) data modelling and significance assessment, and 3) data integration. We identified two sets of enhancer-gene associations (EGAs): the flexible set of ~230K EGAs (*flexibleC)*, and the stringent set of ~60K EGAs (*stringentC*). We additionally identified major differences across previously published computational models of enhancer-gene associations; mostly in the location, number and properties of defined enhancer regions and EGAs. All reg2gene functions were gathered in the reg2gene R package.

We performed detailed benchmarking of seven sets of computationally modelled EGAs, but showed that none of the currently available benchmark datasets could be used as a "golden-standard" benchmark dataset. To account for that observation, we defined an additional benchmark set of positive and negative EGAs with which we showed that the *stringentC* model had the highest positive predictive value (PPV) across all analyzed computational models. We reviewed the influence of EGA sets on the functional analysis of risk SNPs and demonstrated the potential of EGAs to identify gene targets of non-coding SNP-gene associations. Lastly, we performed a functional analysis to detect novel gene targets, enhancer pleiotropy, and mechanisms of enhancer activity. Altogether, this work advances our understanding of enhancer-mediated gene expression regulation in health and disease.

KEYWORDS: gene expression regulation, enhancers, computational modelling, enhancer-gene associations, reg2gene.

Zusammenfassung

Das menschliche Genom enthält Millionen von regulatorischen Elementen - Enhancer -, die die Genexpression quantitativ regulieren. Es wurden zahlreiche experimentelle und rechnerische Ansätze entwickelt, um Enhancer mit ihren Gen-Targets in Verbindung zu bringen. Trotz der enormen Fortschritte im Verständnis, wie Enhancer die Genexpression regulieren, fehlt jedoch immer noch ein systematischer, integrativer und zugänglicher Ansatz zur Entdeckung und Dokumentation von cis-regulierenden Beziehungen im gesamten Genom.

Um dieser Herausforderung zu begegnen, haben wir einen neuartigen rechnergestützten Ansatz entwickelt, der die Genexpression ~ Enhancer-Aktivität (reg2gene) modelliert und integriert. reg2gene wurde auf der Grundlage umfangreicher Datenmodellierung und -integration entwickelt und besteht als solches aus drei Hauptschritten: 1) Datenquantifizierung, 2) Datenmodellierung und Signifikanzbewertung und 3) Datenintegration. Wir identifizierten zwei Sätze von Enhancer-Gene-Assoziationen (EGAs): den flexiblen Satz von ~230K EGAs und den stringenten Satz von ~60K EGAs (über den in drei oder mehr Publikationen berichtet wurde). reg2gene Funktionen wurden im reg2gene R-Paket gesammelt.

Darüber hinaus charakterisierten wir die Unterschiede zwischen zuvor veröffentlichten Berechnungsmodellen von Enhancer-Gen-Assoziationen und zeigten, dass sie sich enorm unterscheiden; in der Lage, Anzahl und Eigenschaften von definierten Enhancer-Regionen und EGAs. Schließlich führten wir ein gründliches Benchmarking von sieben rechnerisch modellierten Sätzen von EGAs durch und überprüften ihren Einfluss auf die funktionelle Analyse von Risiko-SNPs. Wir zeigen, dass keiner der verwendeten Benchmark-Datensätze als die "goldene" Standardmethode zum Testen der Leistung von Berechnungsmodellen angesehen werden kann. Um dieser Beobachtung Rechnung zu tragen, definierten wir einen zusätzlichen Benchmark-Satz positiver und negativer EGAs und zeigten, dass das *stringentC*-Modell über alle analysierten Rechenmodelle hinweg den höchsten positiven Vorhersagewert (PPV) aufwies. Schließlich demonstrierten wir das Potenzial von EGAs zur Identifizierung von Gen-Targets nicht-kodierender SNP-Gen-Assoziationen und führten eine funktionelle bioinformatische Analyse der zugrunde liegenden Enhancer-Regionen durch, um neue Gen-Targets, Enhancer-Pleiotropie und Mechanismen der Enhancer-Aktivität zu erkennen. Insgesamt fördert diese Arbeit unser Verständnis der Enhancer-vermittelten Genexpressionsregulation bei Gesundheit und Krankheit.

SCHLÜSSELWÖRTER: Genexpressionsregulierung, Enhancer, Computermodellierung, Enhancer-Gen-Assoziationen, reg2gene.

Acknowledgments

I started my PhD with one expectation - to learn. The PhD title was my hope - not my goal.

I expected to become more competent as a scientist, but I am more human now.

Nowhere in my plans, I imagined such turbulent, personality- and life-changing years of my life.

I lost WE. I achieved ME.
I lost my health. I reached the Top again.
I relearned how to walk and climb. I became the CAC&C4C.
I lost my love for science. I kept doing my PhD.

I made mistakes. I learned to accept them.

BUT!

I was never alone. I actively surrounded myself with great human beings and invested my time to build strong friendships. I was emotionally supported by many incredible individuals...my colleagues, friends and family. It has been amazing to have you around. Thank you!

To my fellow climbers, mountain rescue members, friends: Enna, Gogo, Malik, Margo, Jurica, Suncica, Corko, Cuki, Valentina, Perica, Tamara, Branko, Robert, Stella, Stephan, Lisa, Sophie, Verena, Meik, Annie, Hannes, Fabi, Vex, Olli, Julia, Marius, Matt, Wolfgang, Wolfgang, Laura, Ben, Jack, Ena, Nazar.

To my colleagues: Kasia, Jona, Alex, Bora, Dilmurat, Wolfgang, Robert, Bren, Ricardo, Dan, Madalin, Ella, Eric, Salah, Ana, Benedict, Russ, Ivana;

To my science crew: Sasa, Petra, Lucija, Nikolina, Zrinko, Mateja. Extended by Dunja, Marieke, Pedro, Matija;

To my Croats: Ivan, Matej, Ivona, Mateja, Igor, Dora, Sanja, Pero;

To my giants: Vedran, Ozren, Luka, Kruno, Altuna;

To my family: Igor, Josip, Mirjana;

To Pouya.

Abbreviations

3C	chromatin conformation capture
4C	chromosome conformation capture on chip
5C	carbon-copy chromosome conformation capture
BS-Seq	bisulfite sequencing
CHiA-PET	chromatin interaction analysis by paired-end-tag sequencing
ChIP-Seq	chromatin immunoprecipitation followed by sequencing
DHSs	DNase I hypersensitive sites
DNase-Seq	DNase I hypersensitive sites sequencing
EGAs	enhancer-gene associations
eQTL	expression quantitative trait loci
GWAS	genome-wide association study
HTS	high-throughput sequencing
NGS	next generation sequencing
RNA-Seq	RNA sequencing
SNP	single nucleotide polymorphism
TF	transcription factor

1

Introduction

1.1. Overview

Complex organisms are constituted of a multitude of specialized cell types. With few exceptions, all cells contain the same genetic material; however, genes are expressed at different levels and in different combinations across different cell types. Multicellular organisms use a variety of mechanisms to regulate the amount of gene products in the cell such as gene expression regulation (Levine and Tjian, 2003), alternative splicing (Berget et al., 1977; Chow et al., 1977), post-translational modifications of proteins (Walsh, 2006), chromatin modification and reordering (Whalen et al., 2016). Nonetheless, most regulation is believed to occur at the level of transcription initiation (Levine and Tjian, 2003).

Transcription is regulated by *cis*-regulatory sequences in the genome that recruit a distinct set of *trans* factors. By promoting or inhibiting the production of mRNAs in each cell, *trans* factors finetune the gene expression of a correct subset of genes. Promoters and enhancers are currently the best-characterized *cis*-regulatory sequences (Andersson et al., 2015). Promoters are located nearby the transcription start sites of genes and integrate a total regulatory input into the rate of transcriptional initiation (Lenhard et al., 2012), whereas enhancers are distal elements that can further refine gene expression across cell types and developmental stages (Banerji et al., 1981; Gerster et al., 1986; Blackwood and Kadonaga, 1998). In most eukaryotes, transcription depends on distal enhancers and their physical separation from promoters is thought to provide an additional level of regulation of gene expression (Levine et al., 2014).

Certain aspects of enhancer-mediated gene expression regulation are known, however, to fully understand and appreciate its complexity it is necessary to systematically identify and characterize all regulatory elements in a genome-wide manner and discern their cell-type specific patterns (Hariprakash and Ferrari, 2019). Especially now, when a daunting amount of high-quality genomic data became available. Multiple approaches have been utilized to study enhancer-mediated long-range gene regulation in a genome-wide manner: predictions using information from the eQTL studies (Rockman and Kruglyak, 2006; Gaffney et al., 2012; GTEx Consortium et al., 2017), (3C)-derived techniques (Dekker et al., 2002; Dostie et al., 2006; Fullwood et al., 2009; Lieberman-Aiden et al., 2009; Simonis et al., 2006), and reporter assays or cellular screens (Arnold et al., 2013; Kheradpour et al., 2013; Kwasnieski et al., 2012; Kvon, 2015; Fulco et al., 2019; Gasperini et al., 2019). The fourth approach, computational modelling of *gene expression ~ enhancer activity*, is the main subject of this thesis.

Hereby, I present my work towards documenting *cis*-regulatory relationships across the genome by means of computational modeling. I frame our current understanding of enhancer-gene associations (EGAs) into the knowledge of genetic susceptibility. I review and characterize differences between published sets of EGAs and perform a thorough benchmarking analysis.

I start by summarizing our current ideas and knowledge about enhancer-mediated gene expression regulation. In particular, I discuss general principles of enhancer functions and mechanisms of their specificity and modularity. In addition, I briefly describe the rise of the high-throughput sequencing technologies (HTS) and how they improved our understanding of enhancer biology. I present techniques, computational tools and algorithms used in this thesis. Lastly, I introduce the current ideas of human (disease) genetics, and how the knowledge about enhancer-gene associations can be used to improve our understanding of genetic susceptibility to human diseases.

1.2. Enhancer-mediated regulation of gene expression

Gene expression has been an important topic of scientific research for a long time. It started as a quest to identify distributor elements of the hereditary information and continued by performing an extensive research on how genetic information is communicated, processed and used to maintain functions and behaviors of the cell. The first considerable challenge was to identify and delineate the structure of genes and other functional elements (Kellis et al., 2014). That is exactly what I set off to describe here.

1.2.1. Defining functional DNA elements in the human genome

A distributor element of the hereditary information was not known until Gregor Mendel, in 1866, postulated the notion of the **gene** (Mendel, 1866). Throughout the twentieth century, genes were attributed to the chromosomes (Waldeyer, 1888; Boveri, 1904; Sutton, 1902, 1903) in the genome (Winkler, 1920).

The initial definition of a gene, simply an element of heredity (Mendel, 1866), evolved as the knowledge about its physical nature expanded. Physical nature of it (which dictates phenotypic traits of an organism) was attributed to **deoxyribonucleic acid (DNA)** - a double helical two-chain molecule composed of covalently bound nucleotides (**Figure 1.1.**, Avery et al., 1944; Hershey and Chase, 1952, Franklin and Gosling, 1953; Watson and Crick, 1953). Each nucleotide was identified to be composed of a **nucleobase** (four possible nucleobases are cytosine [C], guanine [G], adenine [A], thymine [T]) and sugar-phosphate backbone (Kossel, 1911) that pair according to Chargaff's rules within a DNA molecule (Chargaff et al., 1952). Along with DNA, RNA molecules were identified to be a genetic information in some viruses (Wagner et al., 1999).

Figure 1.1. The structure for deoxyribose nucleic acid (DNA) suggested by Watson and Crick in 1953. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fiber axis. Figure reprinted from (Watson and Crick, 1953).

Recently, in April 2020, I decided to search the web by googling the term "gene" aiming to identify the most recent definition. Among many definitions, I found the following:

"*Gene is a specific sequence of nucleotides in DNA or RNA that is located usually on a chromosome and that is the functional unit of inheritance controlling the transmission and expression of one or more traits by specifying the structure of a particular polypeptide and especially a protein or controlling the function of other genetic material.*[1]"

.....

Genes are not the only functional elements coded in DNA. The era of annotating **genome**(s) - attaching biological information to sequences - started with the development of DNA sequencing methods (Maxam and Gilbert, 1977; Sanger et al., 1977). The first sequenced gene was the gene

[1] https://www.merriam-webster.com/dictionary/gene

for Bacteriophage MS2 coat protein (Min Jou et al., 1972). Later, the same group of scientists determined the first two genomes - the complete sequence of bacteriophage MS2-RNA (Fiers et al., 1976) and Simian virus 40 (Fiers et al., 1978). In the second half of the twentieth century, many teams focused on sequencing the entire genomes. However, the ultimate goal was always to sequence **the human genome**. Nonetheless, in eukaryotes, genes sequences were found to be interrupted by the elements that do not directly code for proteins: **introns** or intervening sequences; whereas coding or expressed nucleotide sequences were named **exons** (Berget et al., 1977; Chow et al., 1977).

Soon, the vast majority of the genome (98%) was identified not to code for genes. Initially, non-coding regions were considered to represent a sequence of DNA without any biological function - the "junk" DNA (Gregory, 2011). Only relatively recently, the "junk DNA" regions were found to harbor various functional sequences including regulatory elements such as enhancers, promoters or silencers. In addition, sequences that code for long non-coding RNAs were discovered (Kapranov et al., 2007). Today, approximately 20,000 protein coding, 10,000 long non-coding RNA genes and millions of regulatory regions have been identified in the (human) genome (ENCODE Project Consortium, 2004; Harrow et al.).

Gene-proximal promoters and distant enhancers are among the most studied regulatory elements in the human genome (Andersson et al., 2015). Their interactions allow tight spatio-temporal regulation of gene expression in a cell type-specific manner. Although enhancers and promoters share similar characteristics, they are, in general, considered to represent separate classes of regulatory elements (Andersson et al., 2015; Core et al., 2014). **Promoters** overlap with, or are located close to the transcription start sites of genes (TSS) thereby integrating total regulatory input into the rate of transcriptional initiation. The structure of human gene promoters can be quite complex, typically consisting of a core promoter and nearby (proximal) transcriptional regulatory elements (Lenhard et al., 2012). Promoters, work together with other regulatory regions, like enhancers, to regulate all stages of RNAPII transcription from RNAPII recruitment to transcriptional elongation (Smale and Kadonaga, 2003).

Enhancers are the first regulatory DNA elements shown to be involved in differential gene transcription (Gerster et al., 1986). They are generally located up to 1Mb from the transcription start sites of genes (Lettice et al., 2003), and increase gene expression regardless of their position, orientation and distance to the promoter (Banerji et al., 1981). They do not occur at a defined

distance from a TSS (Blackwood and Kadonaga, 1998); do not necessarily regulate the closest gene (Mifsud et al., 2015; Schoenfelder et al., 2015; Javierre et al., 2016), can regulate more than one gene (Gao et al. 2016; Cao et al. 2017), do not have a specific sequence motif or structure for their univocal genome-wide identification (Visel et al., 2009; Roadmap Epigenomics Consortium et al., 2015) and are very cell-type specific (Joshi, 2014).

1.2.2. Identifying non-genetic, sequence-independent elements of gene regulation

In addition to information coded in the sequence of our DNA, many sequence-independent processes can modulate gene expression patterns in a cell (Rivera and Ren, 2013). Proteins, their (chemical) modifications and the 3D genome structure were recognized to have an important role in organization, maintenance and communication of genetic information. Although its definition changed during the years (Holliday, 1990; Waddington, 1959), the word "epigenome" has been generally used to describe sequence-independent processes that modulate gene expression patterns in a cell-type-specific manner thereby "extending" information stored in the genomic sequence (Rivera and Ren, 2013).

Proteins interact regularly with DNA and RNA. At the frontline of DNA-protein interactions is the **chromatin** - a DNA-histone complex that enables packaging of the mammalian two meters long genome (when in its unfolded, linearized form) within the nucleus (Luger et al., 1997). In the process of gene expression, DNA and RNA molecules frequently interact with protein complexes. For example, the process of transcription in humans is mainly performed by **RNA polymerase II** (Pol II; Roeder and Rutter, 1969). Pol II binds DNA and interacts with one or more general transcription factors (protein complexes): TFIIA, TFIIB , TFIID, TFIIE, TFIIF, and TFIIH (Tang et al., 1996). Many other transcription factors, coactivators (histone acetyltransferases p300 and CBP; Ogryzko et al., 1996), modifying enzymes such as DNA and histone methyltransferases (DNMT; Viré et al., 2006; Iyer et al., 2016;), histone deacetylase (Taunton et al., 1996), and chromatin remodelers (Stern et al., 1984; Pazin and Kadonaga, 1997) were found to be implicated in transcription (regulation). In addition, protein complexes such as DNA helicases (Manosas et al., 2010), topoisomerases (Champoux, 2001), etc. were found to interact with DNA/RNA.

The genome and chromatin structures were shown to support gene expression and its regulation. DNA interacts directly with histone proteins to form nucleosome structures ordered in 10-nm

fibers (repetitive nucleosome motifs of ~200 bp) and organized into 30-nanometer chromatin fibres (Olins and Olins 1974; Kornberg 1974; Woodcock et al. 1976; Finch and Klug 1976; Davey et al. 2002). Early experiments on eukaryotic chromatin compaction identified differences in the degree of compaction between genomic regions containing expressed genes with transcriptionally silent regions (Axel et al., 1973). Meanwhile, chromatin was identified to be a dynamic structure that allows and restricts transcription factor binding (Bell et al., 2011), whereas **DNA accessible regions** (that correspond to the regions depleted of nucleosomes) were associated with transcriptional activity and active enhancers and promoters (Gross and Garrard, 1988).

Recently, spatial arrangements of the chromatin and nucleus, such as active chromatin hubs (Tolhuis et al., 2002), lamina-associated domains - LADs (Guelen et al., 2008), topologically associating domains - TADs (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012), chromosome compartments and territories (Naumova et al., 2013), were identified. Chromatin interactions with elements of cohesin complex (Sofueva et al., 2013), CTCF proteins (Zuin et al., 2014), and some additional factors (de Wit et al., 2013; Beagan et al., 2017) were suggested to underlie a higher-ordered spatial arrangement of the chromatin and nucleus. Importantly, topologically associating domains (**TADs**) are generally defined as regions in the genome characterized by a high level of chromatin interactions occurring within them, some of which can be enhancer-gene interactions (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012; Lupiáñez et al., 2015). They correspond to the 3D genome organization structures that are approximately 100 kb to 1 Mbp in size, are identified in a wide range of metazoans and show striking conservation across species and cell types (Dixon et al., 2012). Although, it was generally accepted that the structural organization of the genome into TADs was maintained across different cell types, recent studies suggested that the genome exhibits significant cell-cell variability in its 3D organization and dynamic chromatin topology (Finn et al., 2019).

.....

Half a century ago, scientists initiated studies of the possible role of chromatin in the regulation of gene expression because they suspected that histones might be repressing transcription (Verdin and Ott, 2015). Soon, acetylation of histones was shown to lower their ability to inhibit RNA synthesis, and the hypothesis that the reversible post-translational histone acetylation is a "dynamic and reversible mechanism for activation and repression of RNA synthesis" was born (Allfrey et al., 1964). Although the first studied epigenomic modification was C5-methylcytosine

(5mC) - a methylated form of cytosine (Hotchkiss, 1948), it was later associated with promoter regions and predicting cell type-specific enhancer activity (Wiench et al., 2011). Meanwhile, **reversible chemical modifications** of proteins, DNA and RNA were identified (**Figure 1.2.**) and multiple **chromatin modifications** were shown to be implicated in gene regulation, maintenance of cellular identity and cell differentiation (Clark et al., 2016).

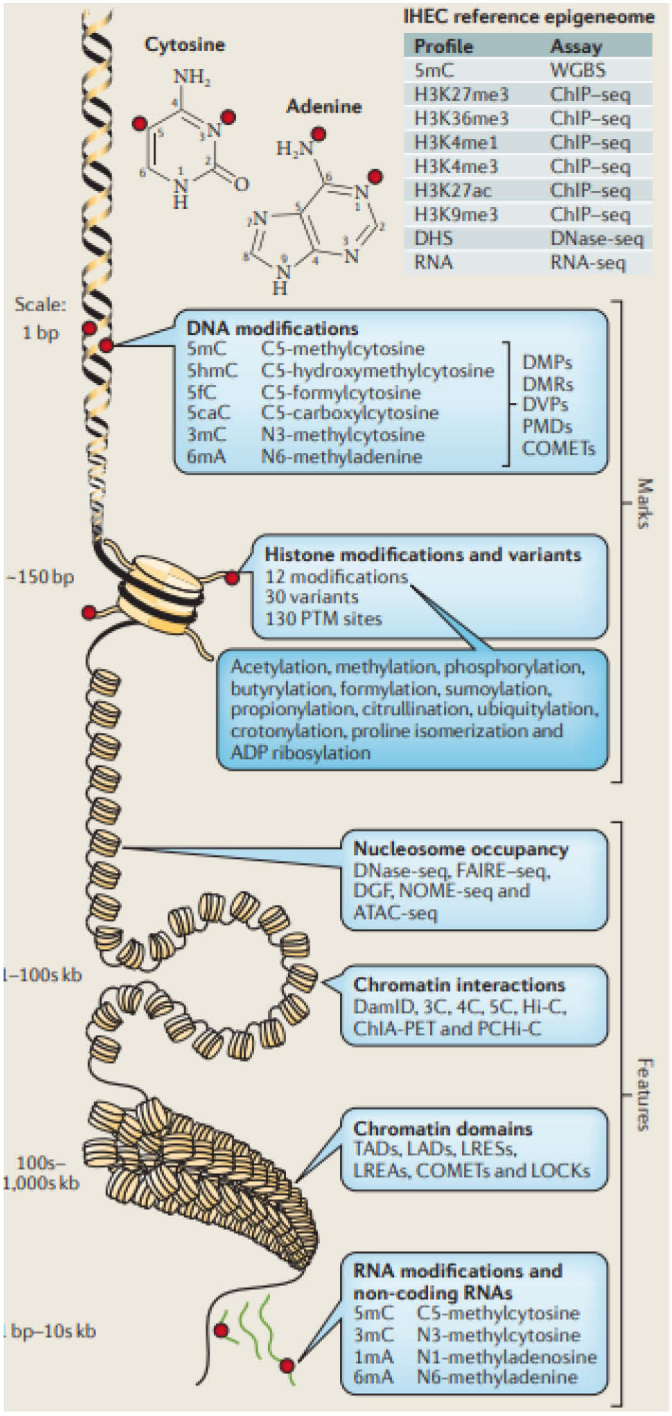


Figure 1.2. A diversity of chemical modifications of DNA, RNA, histone proteins and the genome structure. Figure reprinted from (Stricker et al., 2017)

Along with the genome structure, chromatin modifications are considered to represent a group of non-genetic factors that influence traits and phenotypes, ageing (Benayoun et al., 2015) and cancer (Berdasco and Esteller, 2010). Specifically, modifications of DNA and histone proteins are thought to convey and retain regulatory information through DNA replication, when most transcription factors dissociate from their binding sites (Jenuwein and Allis, 2001).

Until recently, six different known epigenomic modifications were recognized at the DNA level; 12 currently known chemical modifications at the histone level (at more than 130 post-translational modifications sites on five canonical and some 30 histone variants), and more than 100 different known RNA modifications (Stricker et al., 2017). However, the number and complexity of newly discovered chromatin marks increases every day as the high-throughput sequencing (HTS) techniques of epigenomic profiling improve. However, only some of them were later shown to affect gene expression without altering the DNA sequence (Gibney and Nolan, 2010).

.....

In summary, I introduced you to the most important elements of the human genome and its main players (subjects and objects of gene expression regulation): DNA, genes, histones and other proteins, chromatin, chemical modifications, the genome structure, etc.
Next, I continue with adding verbs in this grammar of gene expression regulation.

1.2.3. Early research of gene expression regulation

Gene expression refers to the multi-step processes of transcription, translation, protein localization, and post-transcriptional modifications (Alberts et al. n.d.). As early as the nineteen-seventies, the conceptual paradigm for understanding transcriptional regulation was established: transcription factors bind promoter sequences to promote (or inhibit) binding of the necessary factors for transcription.

The first leap in our understanding mechanisms of gene expression was achieved when the notion of a gene was associated with production of a single enzyme that, in turn, affects a single step in a metabolic pathway - the famously titled "***one gene – one enzyme***" hypothesis (Beadle and Tatum, 1941). The following experiments confirmed the flow of information from DNA to protein through the RNA molecule (Brenner et al., 1961; Crick et al., 1961; Leder and Nirenberg, 1964; Nirenberg and Matthaei, 1961). This was later reformulated as the Central Dogma of biology:

"*Genetic information is transcribed from DNA to RNA and then translated from RNA into protein."* (Crick, 1958, 1970).

Around the same time, the first experiment on gene regulation was done in bacteria (Jacob and Monod, 1961; Pardee et al., 1959). It identified stretches of DNA that act as regulator genes, repressors, co-repressors and revealed certain molecular mechanisms of gene expression. Later, the two-signal regulation system of operons was discovered in *E.coli* (Ptashne et al., 1976). It was exemplified with the *lac* operon where the same regulatory protein acts either as a repressor or an activator depending on the position near or within the promoter (Malan and McClure, 1984). Since in prokaryotes, gene control serves mainly to enable them to adapt to changes in the environment and optimize growth and cell divisions, their regulation was proven to be too simple for eukaryotes. In eukaryotes processes of cell differentiation, morphogenesis, and accurate response to environmental stimuli require precise regulation.

Many additional mechanisms of gene regulation were identified in eukaryotic gene expression such as alternative splicing (Berget et al., 1977; Chow et al., 1977), post-translational modifications of proteins (Walsh, 2006), chromatin modification and reordering (Whalen et al., 2016), gene expression regulation (Levine and Tjian, 2003), etc. However, most regulation in eukaryotes is believed to occur at the level of transcription initiation (Levine and Tjian, 2003). This is later supported by the observation that the RNA content differs greatly for different cell types and it correlates with protein abundance (Schwanhäusser et al., 2011). In addition, transcriptional mis-regulation is associated with diseases such as cancer (Dawson and Kouzarides, 2012).

Transcription is initiated at core-promoters that recruit RNA polymerase II and together with general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) assemble the pre-initiation complex (PIC) to finetune the accurate position of initiation and direction of transcription (Roeder,

1996). However, core-promoters cannot support efficient transcription on their own and they commonly exhibit only low basal activities (Lenhard et al., 2012). Thus, cell-type-specific activation of transcription is thought to be mainly regulated by enhancers (Levine et al., 2014).

1.2.4. General principles of enhancer function

"**Exactly how do enhancers modulate gene expression?**" represents one of the central questions of genomics.

Decades of studies revealed that in order to activate (or repress) transcription in time and cell-type specific manner enhancers recruit specific TFs and mobilize cofactors such as p300 (Visel et al., 2009) or Mediator (Kagey et al., 2010; Zuin et al., 2014), bind Pol II and general transcription factors and assemble the pre-initiation complex (PIC; **Figure 1.3.,** Roeder, 1996) via binding to short recognition sequences: transcription factor (TF) binding sites. Cofactors do not typically bind DNA directly whereas they are bound by different TFs via protein-protein interactions. They execute various biochemical activities - predominately modifying and remodeling the chromatin (Visel et al., 2009b; Kagey et al., 2010; Zuin et al., 2014) which ultimately leads to precise activation or repression of transcription (Shlyueva et al., 2014). Thus, enhancers are considered to be the workhorses behind the transcriptional activation.

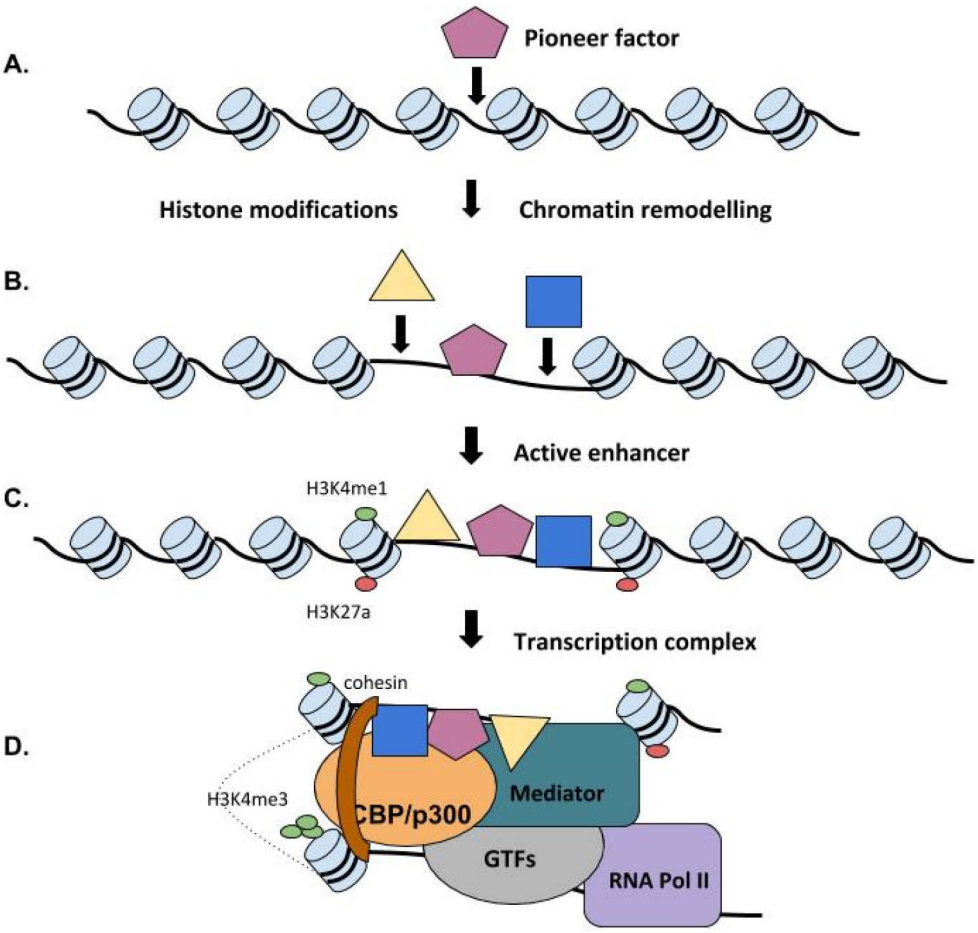


Figure 1.3. Schematic representation of enhancer activation. In the first stage, (A) enhancer is inactive, packaged into chromatin and poorly accessible to the DNA-binding transcription factors. (B) Schematic view of chromatin being opened by a 'pioneer factor' - a DNA-binding protein that is able to occupy inaccessible regions and attract chromatin remodeling proteins and histone modifying enzymes that (C) deposit enhancer-typical histone modifications such as H3K4me1 and H3K27ac and make enhancer accessible to other TFs. (D) Representation of enhancer-promoter interaction via DNA looping: enhancer-binding and promoter-binding proteins and complexes are indicated; such as specific and general transcription factors (TFIIA, TFIIB , TFIID, TFIIE, TFIIF, and TFIIH; here denotes as GTFs), cohesin and the Mediator - a multiprotein complex that can stabilize enhancer-promoter interaction, coactivators such as p300 (an enzyme that has histone acetylase activity) and RNA Pol II.

On the other hand, it was considered that enhancers act by increasing the RNA polymerase density (number of RNA polymerase II molecules) within the linked gene (Weber and Schaffner, 1985; Weber and Schaffner, 1985; Gerster et al., 1986) or facilitate the recruitment of the pre-initiation complex (PIC) to the target promoter and thereby activate transcription (Szutorisz et al., 2005). Since neither, the recruitment of polymerase to promoters is necessarily sufficient to activate transcription (Dorris and Struhl, 2000), nor the polymerase binding was shown to represent a key control point in transcription in mammals (Bartman et al., 2019) the current view is that enhancers sometimes act through regulating the **release of polymerase** from promoter-proximal pausing

after it transcribed approximately 40 nucleotides (Muse et al., 2007; Zeitlinger et al., 2007; Core et al., 2008). Nevertheless, this is still a subject of extensive research.

Importantly, enhancers were found to function when in **physical proximity** to a target promoter (Müeller-Storm et al., 1989). It has been suspected that an underlying physical interaction is facilitated by **DNA looping** - a mechanism by which enhancers and promoters are brought together in the 3D space of a nucleus (Su et al., 1991; Schleif, 1992). Later, DNA looping was shown to, at least partially, facilitate precise regulation of gene expression (Tolhuis et al. 2002; Vernimmen et al. 2007; Deng et al. 2012; Jin et al. 2013) and recent single molecule imaging approaches confirmed that enhancer-promoter contact is necessary for transcriptional activation (Chen et al., 2018a). It is important to note that looping interactions between enhancers and promoters are not necessarily stable interactions, rather form and dissociate dynamically (Chen et al., 2018a).

1.2.5. Hierarchical mechanisms of enhancer specificity

"**How activating regulatory information is communicated from enhancers to their correct promoters?**" is another important question of modern biology.

Historically, many examples of gene expression regulation came from researching *Drosophila melanogaster* and suggested that the precise enhancer–core-promoter targeting is orchestrated at several different levels. For example, one of the first studies of *Drosophila melanogaster* gene expression regulations identified three neighboring genes: decapentaplegic (dpp), SLY1 homologous (Slh) and out at first (oaf), and showed that they were not necessarily regulated by the nearest regulatory regions. Specifically, several dpp enhancers were positioned closer to Slh and oaf, but those two genes were remarkably unaffected by the dpp elements (Merli et al., 1996). These observations were confirmed in vertebrates; tight spatiotemporal regulation of Hox genes was shown to be essential for correct patterning of target tissues and regulated by a hierarchy of molecular controls (van der Hoeven et al., 1996; Sharpe et al., 1998). In addition, Tolhuis et al. (2002) showed that the locus control region (LCR) of the mouse beta-globin locus in the fetal liver was in more frequent contact with the beta-globin promoter than with intervening sequences in liver cells where beta-globin is expressed. However, in brain tissue, which does not express beta-

globin, the contact was absent. Subsequent research identified additional regulation at the level of regulatory domains, chromatin modifications, DNA accessibility, and sequence-encoded specificities that engage different regulatory proteins (Zabidi and Stark, 2016).

For example, the disruption of topologically associating domains - **TADs** - was shown to lead to de novo interactions of enhancers and promoters, gene misexpression and/or occurence of diseases (Spielmann et al., 2012; Spielmann and Mundlos, 2013; Ibn-Salem et al., 2014; Lupiáñez et al., 2015), which suggested that the spatial organization of the genome represents one level of gene expression regulation. Enhancers and their target genes have been frequently found to co-localize in the same TAD region (Symmons et al., 2014) and a role of TADs in directing (or restricting) enhancer function in transcriptional regulation is supported by the fact that TAD boundaries are depleted between enhancer-promoter interactions (Kvon et al., 2014). Many evidence supports a role of TADs in restricting or directing enhancer function during transcriptional regulation including the fact that TAD borders hinder the spreading of chromatin marks associated with transcriptional activity (Narendra et al., 2015; Tsujimura et al., 2015). Nevertheless, TADs do not represent the only level of regulation of enhancer function - many genes co-localized in the same TAD region are not co-expressed and enhancers can be selective for certain promoters (Calhoun et al., 2002).

The regulation of **DNA accessibility** and enhancer-promoter tethering are likely to be involved in controlling enhancer – core-promoter targeting as well (Calhoun et al., 2002; Fakhouri et al., 2010). Nucleosomes were shown to represent a barrier for the access of TFs to their target sites *in vivo* and the activation of transcription is correlated with the reorganization of nucleosomes at enhancer elements, e.g. active enhancers are devoid of nucleosomes (Schones et al., 2008, He et al., 2010). **Nucleosome positioning** and DNA accessibility are modified by a dynamic, enzyme-assisted process of chromatin remodelling that ultimately allows access of regulatory transcription machinery proteins to the condensed genomic DNA (Stern et al., 1984; Pazin and Kadonaga, 1997).

Although active enhancers enhancers are commonly “nucleosome-free” regions, the histones in the flanking nucleosomes often carry specific post-translational modifications that likely affect their function and gene expression. **Histone modifications** are considered to act mainly in two ways: either by altering the biophysical properties of chromatin and influencing chromatin compaction or by supporting binding of additional transcriptional cofactors that precisely finetune

gene expression (Winter et al., 2017). Elements with the biochemical signatures of enhancers (e.g. high level of H3K27ac) have highly cell-type specific activity and that supports their role as enhancers (Calo and Wysocka, 2013). In addition, acetylated histones were shown to form a less compact chromatin conformation (Garcia-Ramirez et al., 1995), which can potentially facilitate the ability of Pol II or transcription factors to access their binding sites on the DNA (Workman and Kingston, 1998). Nevertheless, the mechanistic role of histone modifications in enhancer function and gene expression regulation remains unclear. Histone marks such as H3K4me1 or H3K27ac were shown not the be sufficient, necessary or even mechanistically involved in transcription, whereas H3K4me3 cannot maintain transcription when the activating transcription factor is not present (Hödl and Basler, 2012; Pengelly et al., 2013) or is rapidly lost, from a previously active promoter region, when the activating transcription factor is removed (Hathaway et al., 2012). Thus, the question posed a long time ago: "Are histone modifications causally related to gene expression or simply its consequence?" (Calo and Wysocka, 2013) remains valid even today.

The observation that reporter genes under control of different **enhancer – core-promoter combinations** exhibit distinct expression patterns suggested that the both enhancer and core-promoter sequence-encoded specificities are an important determinant of enhancer-targeting (Merli et al., 1996; Ohtsuki et al., 1998; Sharpe et al., 1998). In other words, the core-promoter sequences falling into different functional classes are activated by distinct types of enhancers by means of different cell-type specific TFs or their combinations (Bender et al., 2001; Landry et al., 2009). It is assumed that their communication is mediated by locally high concentration of specific cofactors that interact dynamically and supported by post-transcriptional modifications (Lemon and Tjian, 2000; Kulaeva et al., 2012). Nevertheless, which combinations of TFs and cofactors are able to activate transcription or which TFs, cofactors, histones, histone modifications or PIC components are involved in the implementation process of their biochemical compatibility represent and an exciting topics of the current research (Zabidi and Stark, 2016).

1.2.6. Towards enhancer modularity

With our greater appreciation of the complexity of genomic organization and improved understanding of regulatory mechanisms, it became apparent that *cis*-regulatory elements do not function in isolation. Many genes were shown to be controlled by multiple discrete enhancer sequences with different tissue specificities (Simonet et al. 1991; Schwartz and Olson 1999; Bender et al. 2001; Burch 2005; Abbasi et al. 2007; Landry et al. 2009; Visel et al. 2009). For example, Schwartz and Olson (1999) showed that seven activating regions and three repressor regions surround the Nkx2-5 gene (the earliest known marker for cardiogenesis) and their differential activity and directed expression patterns are used for the tight spatio-temporal regulation of Nkx2-5: differential expression patterns demarcated distinct subpopulations of cardiomyocytes within cardiac compartments, but none of the enhancers could individually account for the time-wise homogenous, but precisely spatially localized, Nkx2-5 expression patterns. Likewise, GATA genes (GATA1-6) are regulated in a modular fashion by sets of enhancers that govern distinct temporal and/or spatial patterns of the overall expression in heart, gut, hematopoietic lineage development and other tissues (Burch 2005).

Strikingly, 65% of the genes involved in mesoderm development of *Drosophila* were found not to be regulated by only a single enhancer. The majority of these genes are regulated by three to five redundant enhancers, of which one or more can be deleted without significant phenotypic effects (Cannavò et al., 2016). For example, the functional deletion studies revealed that the loss of one of the two redundant enhancers in *Drosophila,* that regulated the expression of the developmental genes *shavenbaby* and *snail*, did not result in a loss of function or phenotype under standardized laboratory conditions but did revealed their importance when the genetically modified embryos were treated with high-temperature stress (Dunipace et al., 2011; Frankel et al., 2010). In addition, redundant enhancers have been generally found to be characterized by a tendency to cluster together, have overlapping expression patterns in reporter assays and their similarities in TF binding sites (Hong et al., 2008). Enhancer redundancy is also a characteristic feature of mammalian genomes - Hay et al., 2016 showed that at the α-globin locus individual enhancers of a cluster act independently and in an additive manner. However, how complex patterns and precision of gene expression during development are achieved and why this involves elements with apparently redundant function remains elusive. For some cases, it is considered

that the key lineage genes in a given cell type tend to be associated with dense cluster(s) of highly active enhancers, often referred to as super-enhancers (Hnisz et al., 2013; Whyte et al., 2013).

.....

In this chapter, I summarized many known (or mentioned unknown) aspects of enhancer-mediated gene expression regulation. However in order to fully understand the complexity of gene expression regulation in the cell, and how it has been changed in respect to the developmental and environmental stimuli, it is necessary to identify all enhancer elements **in the genome**, discern their cell-type specific patterns and ultimately associate them with their target genes.

To achieve that we needed to witness the rise of the high-throughput sequencing (HTS) technologies. Their application in the genome-wide identification and characterization of enhancer regions is the main subject of the following chapter.

1.3. Genome-wide characterization of enhancers in the era of HTS technologies

Genome-wide techniques that produce descriptive data about the chromatin landscape of the genome are collectively called epigenomic profiling technologies (Stricker et al., 2017). They derive epigenome profiles that were shown to provide a proxy information useful to facilitate annotations of the human genome and identify functional sequences in a genome-wide, comprehensive and potentially unbiased manner. In general, epigenomic technologies stand upon the high-throughput sequencing technologies, that, due to their major shift in assay capacity as compared to the low-throughput technologies, allowed questioning of the previous locus-centered conclusions and allowed researchers to extrapolate the previous findings to other parts of the genome (Rivera and Ren, 2013).

In this section, I focus on the recent technological advances and how they are reflected on the identification and characterization of enhancer regions.

1.3.1. The historic expansion of HTS technologies

Although the main goal of the Human Genome Project (HGP) was mainly to obtain the sequence of the human genome (McPherson et al., 2001; Venter et al., 2001; Collins et al., 2003), this project initiated major technological improvements. Sanger sequencing was replaced by the Next-Generation Sequencing (NGS) and the rise of the high-throughput sequencing (HTS) technologies was initiated; the growing array of sequencing assays were developed (Margulies et al., 2005; Ronaghi et al., 1996).

Today, HTS applications exceed far beyond simply sequencing genomes; sequencing assays were developed to characterize the transcriptome, map the regulatory information and three-dimensional organization of the genome, sequence microbiomes, etc. (**Figure 1.4.**, Reuter et al., 2015). For example, one of the first high-throughput sequencing assays developed was DNase-Seq (Crawford et al., 2006). It maps "open" regions of the genome using DNase I digestion followed by

DNA sequencing. Around the same time, ChIP-Seq - a method to map transcription factor (TF) binding sites and identify chromatin modification was developed (Johnson et al., 2007), as well as the RNA-Seq - a method to characterize transcriptome (Nagalakshmi et al., 2008). The HiC - a method that allowed the first unbiased, genome-wide research of chromatin organization was announced in 2009 (Lieberman-Aiden et al., 2009).

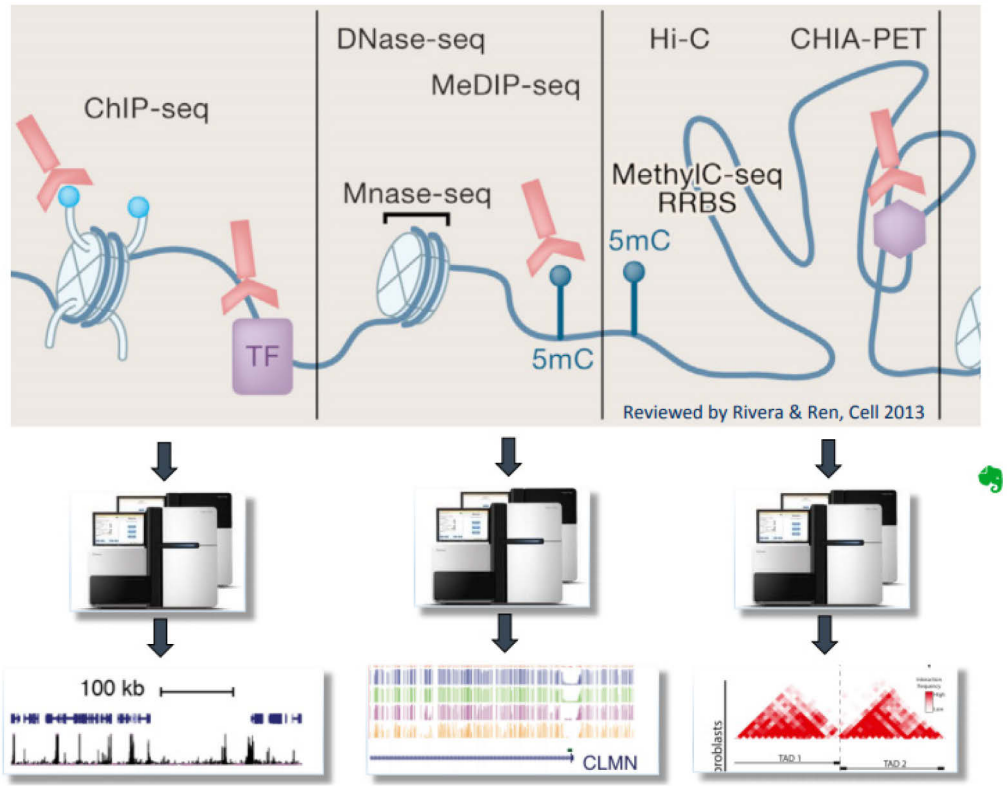


Figure 1.4. Some of the sequencing-based technologies for mapping human epigenomes. ChIP-Seq is a high-throughput method to probe DNA-protein associations (Johnson et al., 2007). It can be used to identify chromatin modification of TF binding sites. DNase-Seq, Mnase-seq, and Hi-C can be used for mapping of chromatin structures. HiC facilitates genome-wide research of chromatin organization (Lieberman-Aiden et al., 2009), whereas DNase-Seq probes chromatin accessibility (Crawford et al., 2006) and Mnase-seq nucleosome positioning (Schones et al., 2008). DNA methylation at 5mC can be probed genome-wide by MethylC-seq or BS-Seq (Cokus et al., 2008). ChIA-PET (Fullwood et al., 2009) is used to profile long-distance DNA interactions mediated by a specific protein. Figure reprinted from (Rivera and Ren, 2013)

Many of the aforementioned HTS technologies were pioneered by The ENCyclopedia Of DNA Elements Project - a follow-up of the HGP aimed to identify all functional elements in the human genome sequence (ENCODE Project Consortium, 2004). As other large-scale consortia-based projects were initiated in the meantime, a colossal amount of genomic and epigenomic data was produced by the HTS technologies and was subsequently used to characterize the human genome

(Bernstein et al., 2010; Harrow et al., 2012; Stunnenberg et al., 2016), study human genetic variation (1000 Genomes Project Consortium et al., 2010), analyze gene expression (GTEx Consortium, 2015), produce three-dimensional maps of mammalian genomes (Dekker et al., 2017), or discover the molecular foundation of human diseases (Cancer Genome Atlas Research Network et al., 2013; Welter et al., 2014), etc.

In this thesis, I analyzed a subset of available high-throughput datasets and data types. I further introduce the basics of technologies which results I used in this thesis: ChIP-Seq, BS-Seq, DNase-Seq, Chromatin conformation capture (3C) and derived techniques, CHiA-PET, RNA-Seq, microarrays and CRISPR-Cas system.

1.3.2. Epigenome mapping technologies

ChIP sequencing

ChIP sequencing or chromatin immunoprecipitation followed by sequencing (ChIP-Seq), is an NGS method developed to study protein-DNA interactions in the nucleus (Gilmour et al., 1986). Although, chromatin immunoprecipitation per se has been in use since 1988 (Solomon et al., 1988), prior pairing with sequencing, it was frequently utilized in combination with hybridization microarrays - in a technique called ChIP-chip, and in tandem with NGS as previously mentioned (Mikkelsen et al., 2007; Robertson et al., 2007; Barski et al., 2007). In short, during the ChIP protocol, DNA and its bound proteins are cross-linked. The DNA-protein complexes are sheared, and protein-specific antibodies are used to select and pull down the DNA fragments associated with the protein of interest. Pulled-down DNA fragments are sequenced and aligned to a reference genome. This enables genome-wide mapping of a protein's activity on the genome at base pair resolution. The ChIP-Seq data represents a signal across genomic coordinates, where each base is associated with a binding strength quantified as a real-valued signal (normalized count of the reads covering each location). Although characterized by a high number of false positives and the presence of artifacts, ChIP-seq still represents the golden standard for mapping genome-wide protein-DNA interactions (Wreczycka et al., 2019).

Bisulfite sequencing

Bisulfite sequencing (BS-Seq) is an NGS method developed to study DNA methylation genome-wide - the methylome. In BS-Seq, DNA sequencing is preceded by bisulfite treatment of the DNA, a reaction which converts cytosine residues to uracil, unless those cytosines are methylated and a comparison with the reference sequence results in identification of the methylated cytosines (Frommer et al., 1992). Bisulfite-treated DNA sequencing is a golden standard for probing the methylome genome-wide (Wreczycka et al., 2017). As previously mentioned, C5-methylcytosine (5mC) was the first studied epigenomic modification (Hotchkiss, 1948) and was shown to change the activity of a DNA segment without changing the sequence especially when located in promoter (enhancer) regions (Wiench et al., 2011).

DNase-Seq

DNase-Seq (DNase I hypersensitive sites sequencing) is an NGS method developed to study chromatin accessibility in a sample. It identifies regions sensitive to cleavage by DNase I enzyme genome-wide - DHSs or DNase I hypersensitive sites (Crawford et al., 2006; Madrigal and Krajewski, 2012). The accessibility of a genomic segment containing genes is a requirement for polymerase to be able to transcribe any genes from that segment, thus chromatin accessibility is an important epigenetic mark that indicates active regions in the genome. It was succeeded by the FAIRE-Seq (Giresi et al., 2007); however, ATAC-Seq has been more commonly used recently (Buenrostro et al., 2013). Briefly, it is performed by first treating DNA with a DNase I enzyme, which cuts loose accessible chunks of DNA and then primers are ligated to the resulting DNA fragments. Primers are amplified and sequenced, and when mapped to the reference genome, they indicate which parts of the genome are accessible (Buenrostro et al., 2013).

Chromatin conformation capture (3C) and derived techniques

Chromatin conformation capture (3C, Dekker et al., 2002) and derived techniques such as chromosome conformation capture on chip - 4C (Simonis et al., 2006), carbon-copy chromosome conformation capture - 5C (Dostie et al., 2006) and HiC (Lieberman-Aiden et

al., 2009) are ligation-based (high-throughput) techniques that have been used to measure chromatin contact frequencies in the nucleus and study the spatial organization of chromatin in a cell. In general, the 3C-based methods consist of several steps including crosslinking with formaldehyde, genome fragmentation using restriction enzymes, ligation of interacting regions and their quantification (Pombo and Dillon, 2015). The main difference between methods is their scope: for example, 3C methods quantify interactions between a single pair of genomic loci (Dekker et al., 2002), whereas HiC is a fully high-throughput method (Lieberman-Aiden et al., 2009). 3C-based methods can be combined with other technologies to adapt them for specific problems, for example, CHiA-PET (chromatin interaction analysis by paired-end-tag sequencing) combines Hi-C and ChIP-seq to detect interactions mediated by a protein of interest (Fullwood et al., 2009).

RNA sequencing

RNA sequencing (RNA-Seq) is an NGS method developed to quantify the level of all RNA transcripts in a sample. It is performed by sequencing cDNA reverse-transcribed from RNA extracted from a sample. Prior to the development of NGS, hybridization arrays were used to measure gene expression, or the relative abundance of different mRNA transcripts in cells (DeRisi et al., 1996). NGS was first used to quantify gene expression by sequencing reverse-transcribed cDNA in 2008 (Morin et al., 2008). Today, RNA sequencing has mostly replaced microarrays for gene expression quantification, thanks to its numerous advantages: less susceptible to cross-hybridization mistakes, has a better dynamic range, offers better detection of highly and lowly expressed genes, provides superior fidelity and it does not require the transcript sequence to be known a-priori (Grabherr et al., 2011; Zhao et al., 2014). In RNA-Seq, reads from the sequencing experiment are aligned to a reference sequence, and each is assigned to the gene (or transcript) it has originated from. The final product is a read count for each transcript.

Microarrays

Microarray is the platform for genotyping - assessing known markers in the human genome - which enables researchers to identify single nucleotide polymorphisms (SNPs) or larger structural changes among millions of markers. Specifically, DNA microarray (DNA

chip or biochip) is a collection of microscopic DNA spots attached to a solid surface - probes - that are used to hybridize a cDNA or cRNA. Probe-target hybridization is then detected and quantified by detection of fluorophore- or chemiluminescence (Taub et al., 1983). Many different DNA microarrays have been developed, especially by Illumina, and used in various genome-wide association studies or eQTL studies such as the GTEx Project (GTEx Consortium, 2015).

CRISPR–Cas system

The type II CRISPR/Cas9 system (Cas9) is a versatile technology for genome engineering (Doudna and Charpentier, 2014; Hsu et al., 2014). In a nutshell, CRISPR–Cas system is targeted to a precise genomic location by a guide RNA (gRNA) - a synthetic RNA complementary with a target DNA sequence (Mali et al., 2013) that is fused to a catalytically inactive variant of the bacterial Cas9 nuclease (dCas9). The enzymatic activity of the Cas9 nuclease is usually terminated by mutation in the RuvC and HNH domains and results in the nuclease-null deactivated Cas9 - dCas9 (Jinek et al., 2012). To successfully modify (add or remove) chromatin marks at the target sites a range of chromatin-modifying enzymes or general transcriptional activator and repressor proteins such as p300 (Hilton et al., 2015), LSD1 (Kearns et al., 2015) , DNMT3A (Rivenbark et al., 2012; Siddique et al., 2013), KRAB (Fulco et al. 2016; Fulco et al. 2019; Gasperini et al. 2019) can be attached to the DNA-binding domains. The grounding behind using genome editing techniques, such as CRISPR–Cas system (Canver et al., 2015), for functional characterization of regulatory regions, lies in the fact that the fusion of chromatin-modifying enzymes (or their catalytic domains) and targetable DNA-binding domain can be used to change a single chromatin mark at particular genomic sites thereby repressing or activating enhancers. In addition, the CRISPR–Cas system allows easy generation of targeting constructs and its potential for multiplexing (Stricker et al., 2017). Recently, cell-based genetic screens - a high-throughput approaches based on CRISPR interference (CRISPRi) - have been recently adapted to evaluate candidate regulatory sequences in their native genomic context and globally capture perturbations of gene expression (Canver et al. 2015; Wakabayashi et al. 2016; Fulco et al. 2016; Diao et al. 2017; Gasperini et al. 2017; Fulco et al. 2019; Gasperini et al. 2019).

1.3.3. Annotating *cis*-regulatory elements from chromatin profiles

Enhancers are marked by characteristic chromatin signatures (**Figure 1.5.,** Heintzman et al., 2007), but lack motifs that could be used for their general identification (Lee et al., 2015; Colbran et al., 2017). Thus, functional data (and an extensive use of epigenome mapping technologies mostly through the large epigenomic consortia; ENCODE Project Consortium, 2004, Andersson et al., 2014) has been used to predict enhancers in a genome-wide manner (Hariprakash and Ferrari, 2019).

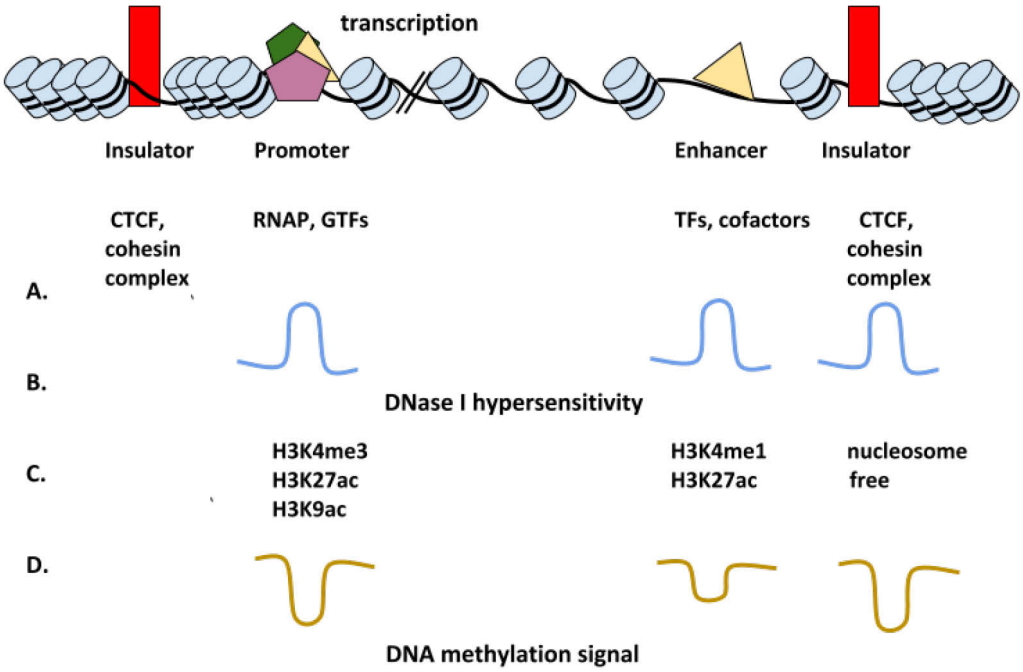


Figure 1.5. A schematic representation of regulatory elements in the genome and their distinctive chromatin signatures that have been commonly used to facilitate their genome-wide identification. Enhancers are, in general, characterized by the presence of: (A) binding of specific TF and cofactors such as p300, (B) chromatin with higher accessibility and are depleted of nucleosomes, (C) high levels of H3K4me1 and H3K27ac and (D) low level of DNA methylation. Distinctive chromatin features for promoters and insulators are indicated as well.

In general, several main approaches were used to predict enhancers: predictions using motifs and conservation (Berman et al. 2002; Kheradpour et al. 2007), chromatin accessibility (Thurman et al. 2012), enhancer–promoter interactions or predictions from transcription factors regulator binding (Visel et al. 2009) and histone modifications (Heintzman et al. 2009; Creyghton et al. 2010).

For example, some methods search for regions in the genome that are highly **conserved across species** (Del Bene et al., 2007; Kheradpour et al., 2007), while others identified genomic regions enriched for transcription factor motif matches (Berman et al., 2002; Kheradpour et al., 2007). ChIP-seq method was frequently utilized to target specific enhancer-related histone marks or identify *in vivo* binding sites for various TFs and cofactors (such as p300; Visel et al., 2009). Later was driven by the rationale that enhancer activity strictly depends on **binding of TFs and cofactors** (Visel et al. 2009; Kagey et al. 2010; Zuin et al. 2014). On the other hand, regions for which the presence of high levels of chromatin marks associated with enhancer activity, such as **H3K4me1 and H3K27ac**, were commonly identified as putative enhancers (Heintzman et al., 2009; Creyghton et al., 2010; Rada-Iglesias et al., 2011). In addition to H3K4me1 and H3K27ac, DNA methylation was found to be a robust predictors of enhancer activity (Varley et al., 2013, Whalen et al., 2016) and its presence/absence was frequently used enhancer predictions (Shlyueva et al., 2014).

As enhancers are generally characterized by chromatin with **higher accessibility** and are **depleted of nucleosomes** (Felsenfeld et al., 1996; Gross and Garrard, 1988), methods that probe chromatin accessibility, such as DNase-seq and ATAC-seq, have been generally used for their genome-wide identification (Thurman et al. 2012). Importantly, the majority of surveyed TFs was found to bind almost exclusively to open chromatin (ENCODE Project Consortium, 2004) and transcription factor binding and DNA accessibility were found to be highly correlated (Kaplan et al., 2011; Pique-Regi et al., 2011).

Lastly, ever since enhancer-derived RNAs were observed at regions marked by enhancer-associated chromatin modifications they have been used for enhancer identification (Kim et al., 2010). Active enhancers produce bi-directional non-coding RNAs (**eRNAs**; Wang et al., 2011), which expression level correlates with the functional activity of the enhancer (Mikhaylichenko et al., 2018) and with the activation of nearby genes (De Santa et al., 2010; Lai et al., 2013). This enabled the use of eRNAs (Andersson et al., 2014) or nascent transcription (Hah et al., 2011) as a marker for annotation of active enhancers across cell lines in a genome-wide manner.

Up today, there is no consensus on which chromatin marks should be used to predict enhancers (Shlyueva et al., 2014; Stricker et al., 2017). To overcome aforementioned non-specificity of individual chromatin marks to predict enhancer regions in the genome, (Ernst et al., 2011) proposed **computational integration** of epigenomic datasets. In a comprehensive epigenomic

study of nine human cell lines, (Ernst et al., 2011) proposed that the human genome could be segmented into regions that carry one of 15 different combinations of chromatin modification marks and that each chromatin state corresponds to a specific category of genomic features. Meanwhile, epigenomics and transcriptomics datasets were integrated computationally using approaches such as CSI-ANN (Firpi et al., 2010), Seqway (Kleftogiannis et al., 2015), ChromHMM (Ernst and Kellis, 2012), DEEP (Hoffman et al., 2012), etc. In addition, separate computational methods that use machine-learning approaches to identify characteristic DNA sequence features in experimentally determined enhancers were developed to use these chromatin features to predict novel enhancers (Kantorovitz et al., 2009; Narlikar et al., 2010).

Nevertheless, only *in vivo* testing of the reporter or *in vivo* editing of the enhancer in transgenic animals have been considered to be a definitive proof of enhancers and their activity (Visel et al., 2007; Catarino and Stark, 2018).

1.3.4. Genome-wide identification of enhancer-gene interactions

Genome-wide approaches did identify numerous putative enhancers, but the vast majority of them have not been functionally tested and we still do not know which genes and in which cell types (if any) enhancers regulate (Arnold et al. 2013; Kvon et al. 2014; Catarino and Stark 2018; Fulco et al. 2019).

Elucidating the function and targets of enhancers remains elusive for multiple reasons: relative location of the enhancer with respect to its target genes can be greatly variable (Lettice et al. 2003), one gene is commonly regulated by more than one enhancer and one enhancer regulates more than one gene (Simonet et al. 1991; Schwartz and Olson 1999; Bender et al. 2001), enhancers are very cell-type specific (Burch 2005; Abbasi et al. 2007; Landry et al. 2009; Visel et al. 2009) and hard to identify genome-wide per se (Hariprakash and Ferrari 2019), or enhancers can “skip” over a proximal to gene to regulate a more distant one (Merli et al. 1996). Remarkably, activation of a specific gene may be activated in multiple cell types by distinct enhancers (Visel et al. 2009).

Multiple approaches have been utilized to study enhancer-mediated long range gene regulation and can be broadly categorized into four categories: predictions using information from the eQTL studies (Rockman and Kruglyak 2006; Gaffney et al. 2012; GTEx Consortium et al. 2017), 3C-related technologies (Dekker et al. 2002; Simonis et al. 2006; Dostie et al. 2006; Lieberman-Aiden et al. 2009; Fullwood et al. 2009), and reporter assays or cellular screens (Arnold et al. 2013; Kwasnieski et al. 2012; Arnold et al. 2013; Kheradpour et al. 2013; Kvon 2015; Gasperini et al. 2019). Lastly, computational modelling has been commonly used to map enhancers to their putative genes (**Figure 1.6.**), however, this will be explained in detail in the next chapter.

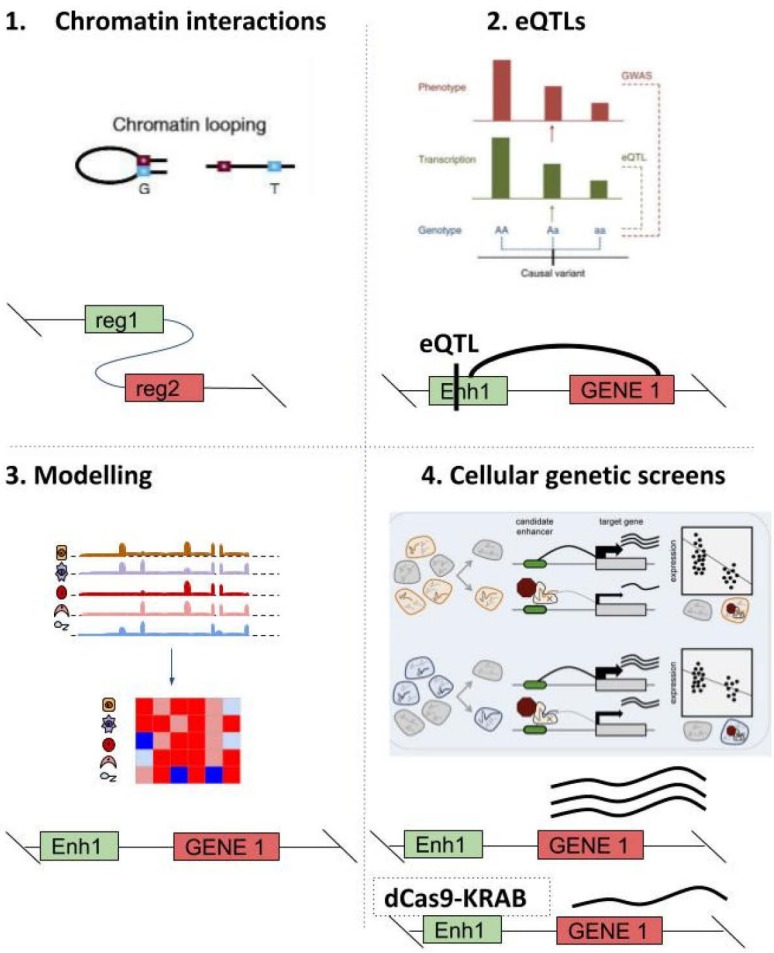


Figure 1.6. Schematic representation of four main methodological approaches used to annotate regulatory regions with their target genes: 1) Information about the 3D genome architecture and chromatin interactions are commonly used as a proxy of enhancer-mediated regulation of gene expression; 2) An overlap between expression quantitative trait locus (eQTL) and regulatory region can be used to assign eQTL associated gene (eGene) to tested enhancer regions. 3) Associations of regulatory regions and their targeted genes have been assessed computationally. 4) A direct functional confirmation or quantitative readout of enhancer activity can be tested in reporter assays and cellular screens.

The fact that enhancers are brought into contact with their target promoters by potentially "looping out" large stretches (up to 1Mb) of the intervening DNA has been exploited to predict enhancers. The development of the chromosome conformation capture (3C) and derived techniques (Dekker et al. 2002; Simonis et al. 2006; Dostie et al. 2006; Lieberman-Aiden et al. 2009; Fullwood et al. 2009) allowed genome-wide assessment of the interaction frequencies between different sites on the same or different chromosomes, which, in turn, provided information about the physical properties of the chromatin fiber and as a proxy for enhancer-mediated regulation of gene expression (Dekker et al. 2002).

An overlap between expression quantitative trait locus (eQTL) and regulatory region can be used to assign eQTL-associated gene (eGene) to tested enhancer regions. eQTLs are a population-based measure of the effect of genetic polymorphism on gene expression (GTEx Consortium et al., 2017). They are defined as nucleotide variants that correlate with differences in gene expression (Rockman and Kruglyak, 2006). Ever since integration of eQTLs with other genomic features was shown to provide a good estimate of the regulatory effect on target genes (Gaffney et al., 2012), they have been frequently used to annotate risk SNPs to their target genes (Welter et al., 2014). On the other hand, if some regulatory region was found to overlap with eQTL this was (could be) used as a proof of a genetic link between the regulatory region and its potential target gene (GTEx Consortium, 2015).

To provide a direct functional confirmation or quantitative readout of enhancer activity, reporter assays (Arnold et al., 2013) and cellular screens were developed (Fulco et al., 2019; Gasperini et al., 2019). *In vivo* based systems enhancer–reporter assays (Kvon, 2015) test regulatory potential of thousands of enhancers using one of defining properties of enhancers - activation of gene transcription from a minimal promoter (Banerji et al., 1981). Reporter transcription can be detected by various methods such as RNA *in situ* hybridization (O'Kane and Gehring, 1987), live imaging of nascent RNA (Bothma et al., 2014), lacZ staining (Kothary et al., 1989) or by activation expression of fluorescent proteins such as GFP (Chiocchetti et al., 1997). The genome-wide versions of enhancer reporter assays - massively parallel reporter assays; MPRAs, (Patwardhan et al., 2009) have been used to test regulatory potential of thousands of enhancers across different cells or tissues simultaneously (Kwasnieski et al., 2012; Arnold et al., 2013; Kheradpour et al., 2013). They are designed in three conceptually different ways: to randomly probe different parts of the genome by creating "enhancer-traps", to directly test enhancers for *in vivo* activity in

enhancer-reporter vectors, or to test enhancers in the context of a longer DNA sequence using, for example, BAC transgenesis (Kvon, 2015).

On the other hand, cell-based genetic screens have been recently adapted to evaluate candidate regulatory sequences in their native genomic context by perturbing sequences using the CRISPRi (dCas9-KRAB) system (Canver et al. 2015; Wakabayashi et al. 2016; Fulco et al. 2016; Diao et al. 2017; Gasperini et al. 2017; Fulco et al. 2019; Gasperini et al. 2019)**.** In a nutshell, chromatin-modifying enzymes or general transcriptional activator and repressor proteins such as p300 (Hilton et al. 2015), LSD1 (Kearns et al. 2015) , DNMT3A (Rivenbark et al. 2012; Siddique et al. 2013), KRAB (Fulco et al. 2016; Fulco et al. 2019; Gasperini et al. 2019) are attached to the CRISPR/Cas9 system to modify chromatin marks at the target sites. Importantly, by using single-cell RNA sequencing (scRNA-seq) screens one can globally capture gene expression perturbations without strong *a priori* hypothesis about target genes (Fulco et al. 2019; Jaitin et al. 2016; Gasperini et al. 2019).

1.4. Functional characterization of enhancers by computational modelling

In genomics and bioinformatics, computational algorithms have been used to support biological research and answer various questions such as analyzing protein-protein interactions (Marcotte et al., 1999; Qi et al., 2006; Papanikolaou et al., 2015), aligning biological sequences (Altschul et al., 1990), annotating enhancers and enhancer-gene interactions genome-wide (Mount and Mount, 2001). Here, I shortly introduce the rationale behind using algorithms to annotate enhancers and associate them with their gene targets. I provide an overview of the currently available approaches.

1.4.1. Computational approaches to identify gene-enhancer associations

Algorithms that were developed to identify enhancer-gene associations can be broadly categorized as: 1) correlation-based; 2) supervised learning-based; 3) regression-based and 4) methods based on other scores (**Table 1.1.**, Hariprakash and Ferrari, 2019). The number and type of features used by an individual algorithm varies, but most of them take into account a combination of gene expression, histone marks, chromatin accessibility and the distance between enhancers and genes (Hariprakash and Ferrari, 2019).

Correlation-based modelling

Correlation corresponds to any statistical relationship, causal or not, between two random variables or bivariate data. It is usually measured by Pearson and Spearman correlation coefficients that range between +1 and −1 (where 1 denotes a total positive linear correlation, 0 no linear correlation, and −1 total negative linear correlation).

As long as enhancers have been recognized as regulatory elements in the eukaryotic genomes, a direct correlation between their activity and transcription rate was recognized (Gerster et al. 1986). However, recently, this property was used to link enhancers with their target genes: for example, the correlation-based enhancer-gene association algorithms, such as PreSTIGE (Corradin et al. 2014) and ELMER (Silva et al.

2019), originate on the rationale that the activity status of an enhancer and its target gene is correlated across multiple cell types Ernst et al. 2011 correlated gene expression with different histone modification marks, including enhancer associated marks H3K27ac and H3K4me1, Sheffield et al. 2013 correlated DNase I hypersensitivity and gene expression, whereas Varley et al. 2013 focused on DNA methylation. Thurman et al. 2012 as well used DNase-seq read coverage at DHS regions and correlated them to DNase-Seq signals at promoters, whereas Corradin et al. 2014 used H3K4me1 ChIP-Seq and RNA-Seq signal to estimate the activity of enhancers and genes, and assess their relationship. Importantly, to identify putative EGAs by correlation methods a certain degree of variance in enhancer activity is required.

Supervised learning-based methods

In general, supervised learning is a type of machine learning that aims to identify a function which maps an input to an output based on input-output pairs. In the case of enhancer-gene associations, supervised learning algorithms leverage a set of EGAs that are assumed to represent “known” true positives and negatives, and build a model that, given the training set, identifies patterns in functional data. Models typically incorporate features derived from epigenomics and transcriptomics data across different cell types in the training set. However, as the number and type of possible features are greatly variable, predictions tend to be very algorithm-specific (Hariprakash and Ferrari 2019). Trained models can be generally used to predict EGAs in other cell types; however, given the very cell type-specificity of enhancers, the reliability of models varies greatly for different cell types (Cao et al. 2017). The main limitation of supervised learning is that the training set requires a set of known positive and negative interactions.

For example, Gao et al. (2016) identified EnhancerAtlas enhancer - promoter associations across 48 cell lines and 22 tissues using IM-PET algorithm - a random forest classifier developed by He et al. (2014) that integrates four types of models/data sources to predict enhancers: TF and promoter activity correlation, correlation of enhancer and promoter activity, enhancer and promoter sequence co-evolution and distance between enhancer and promoter. On the other hand, enhancer-to-promoter distance, H3K4me1, H3K4me3,

and H3K27ac ChIP-Seq and RNA-Seq data was used as an IM-PET input to predict enhancer-promoter associations.

Regression-based algorithms

Regression-based methods work on the rationale that, since multiple enhancers can regulate a single gene, a combinatorial rather than pairwise approach should be used to predict enhancer-gene associations (Reuter et al., 2015). Regression-based methods have the ability to determine the relative influence of one or more predictor variables and thus can assess significant relationships and at the same time assess the strength of impact for multiple variables (Hariprakash and Ferrari, 2019). The main limitation of regression methods is that they rely on arbitrarily chosen parameters, for example a maximum number of enhancers tested around each TSS. In addition, they do prefer larger datasets - number of cell types with available functional data used to build the models (Hariprakash and Ferrari, 2019).

One of the regression based algorithms that I analyzed in this thesis is **FOCS (FDR-corrected OLS with Cross-validation and Shrinkage)**. In FOCS, Hait et al. (2018) used ordinary least squares regression method and across-cell type epigenomic signals to learn predictive models. They modelled activity of promoters based on the activity level of its ten closest enhancers (within a window of ±500 kb around the gene's TSS) using DHS signals from the ENCODE (106 cell types), Roadmap DHS (73 different cell types and tissues), FANTOM5 CAGE (600 human cell lines and primary cells), and GRO-Seq expression data (23 different human cell lines).

On the other hand, **JEME** (Cao et al., 2017) is a hybrid method that combines a regression step followed by a random forest classifier (supervised learning) trained to predict EGAss based on each cell type-specific data. First, they built per-gene regularized regression models (LASSO and elastic net) of gene expression and enhancer activity using across-cell-type signals: RNA-Seq, DHS, H3K4me1, H3K4me3, and H3K27ac ChIP-Seq and FANTOM5 CAGE datasets. In the second step, they assessed cell-type specific enhancers and individually, for each cell-type, they built a Random Forest classifier based on the previously inferred information (prediction errors, distance between enhancers and TSS, and levels of activity signals for analyzed chromatin marks) and trained it with the 'gold-

standard' answers: ChIA-PET, Hi-C or eQTL reported interactions. They assessed model performance using within-cell type or across cell-types cross-validation.

Table 1.1. Pros and cons of four commonly used computational approaches that model enhancer-gene associations.

	Pros	Pros	Example method
Correlation	● Identify multiple targets of an enhancer and can directly derive a quantitative measure of the strength of association	● confounding correlation patterns in case enhancer regions are defined at a resolution higher than that of the functional chromatin mark data used to measure their activity ● availability of genomic data over a large panel of cells, with comparable quality and resolution across all conditions ● is not a complete consensus about which epigenomic or transcriptomic data assess enhancers activity the best ● confounded by enhancers active only in one or few cell types ● do not directly consider the fact that multiple enhancers can act on a gene in a cooperative fashion	● Varley et al. 2013 ● Thurman et al, 2012 ● Corradin et al. 2014 ● Ernst et al. 2011 ● Sheffield et al. 2013
Supervised ML	● once the model is trained, it can predict EGAs in other cell types	● reliability of the model can vary greatly when applied to different cell types ● requires a set of known positive as well as negative interactions	● JEME (Cao et al. 2017); ● FOCS (Hait et al., 2018)
Score based	● More flexible prioritization of EGAs by adjusting a single threshold on the score.	● rely on a number of assumptions and arbitrarily defined parameters or weights	● GeneHancer (Fishilevich et al., 2017);

Other methods: score-based and data integrations

In general, algorithms can combine experimental results into a single score that assess the modelling success. For example, GeneHancer (Fishilevich et al., 2017) quantifies a single quantitative score that defines the strength of association between enhancers and target genes, taking into account multiple types of information. GeneHancer, for example, combines eQTLS, TF-target gene co-expression, eRNAs, capture Hi-C and genomic distance between enhancer and target gene by performing various data transformations and weights to combine aforementioned heterogeneous datasets into a single quantitative value. The main limitation of such approaches is that they rely on a number of assumptions and arbitrarily defined parameters or weights (Hariprakash and Ferrari, 2019).

As in the case of enhancer definition, to overcome inherent weaknesses of different high-throughput technologies which results are used to assess links between enhancers and their target genes, data integration of different epigenetic datatypes was proposed (Ernst et al. 2011). Computational approaches such as PreSTIGE (Corradin et al. 2014), TargetFinder (Whalen et al. 2016), SPEID (Singh et al. 2019), RIPPLE (Roy et al. 2015), IM-PET (He et al. 2014), DECRES (Li et al. 2018) were developed to integrate information about the 3D genome architecture, eQTLs, levels of epigenetic marks, and/or results of reported assays. For example, with TargetFinder, Whalen et al. 2016 integrates chromatin states with Hi-C interaction maps to predict individual and cell type-specific enhancer–promoter interactions using. However, aforementioned methods run for only a handful of cell lines. Some o the aforementioned methods belong to this category as well: EnhancerAtlas (Gao et al., 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al., 2017), FOCS (Hait et al., 2018), HANCER (Wang et al. 2019). However, their predictions were never systematically analyzed, nor were their differences characterized.

1.5. Statistical background of the reg2gene algorithm

1.5.1. Algorithms implemented in reg2gene

As we, later on, implemented five algorithms in reg2gene to account for different relationships between enhancer activity and gene expression: Pearson and Spearman correlation, distance correlation, elastic net and random forest, I will shortly introduce their statistical foundations.

Elastic net

We implemented an elastic net algorithm (Zou and Hastie, 2005) - a penalized regression method that was shown to work well for high-dimensional data with few examples - "large p, small n" cases and in context of association studies (Waldmann et al., 2013). Along with lasso (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970; Tikhonov et al., 1995), elastic net simultaneously assesses the strength of an impact of enhancers (assess the errors terms in predicting TSS activity based on the activity of all candidate enhancers) on their target genes and identifies significant models (Friedman et al., 2009).

Elastic net is implemented in the glmnet R package (Friedman et al., 2009) that fits a generalized linear model via penalized maximum likelihood.

The glmnet solves the following problem:

$$\min_{\beta 0,\beta} \frac{1}{N} \sum_{i=0}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda\left[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1\right]$$

over a grid of values of regularization parameter lambda λ. l(y,η) is the negative log-likelihood contribution for observation i. The elastic-net penalty is controlled by α and to actually execute the elastic net algorithm, we set α to be 0.5 (α=1 corresponds to lasso, whereas α=0 corresponds to ridge regression). The ridge penalty shrinks the coefficients of correlated predictors towards each other while the lasso tends to pick one of them and discard the others. On the other hand, the elastic-net penalty mixes these two methods, by linearly combining the L_1 and L_2 penalties, and if predictors are correlated in groups, it

tends to select the groups in or out together. As compared with least square regression, elastic net extends the least squares minimization with a term that includes the values of the predictor coefficients, β, in the minimization process. Thus, elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together (normal regression would assign all weight to just one of those sets).

Random forest

Random forest is a machine learning algorithm that uses an ensemble of decision trees (Breiman et al., 1984; Friedman et al. 2001) to solve classification, regression or other computational problems (Breiman, 2001).

It operates by constructing a multitude of decision trees at training time and returning the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees as an output (Ho, 1995). Each of the classification trees is built using a bootstrap sample of the data, and at each split the candidate set of variables is a random subset of the variables. Thus, random forest uses both bootstrap aggregation (Friedman et al. 2001), and random variable selection for tree building. Each tree is unpruned (grown fully) and the algorithm seeks an ensemble that can achieve both low bias and low variance (from averaging over a large ensemble of low-bias, high-variance but low correlation trees). The most important parameters to choose in RF are the number of input variables tried at each split, number of trees to grow for each forest and the minimum size of the terminal nodes (Liaw and Wiener, 2002).

Due to its characteristics, especially the ability to return the measures of variable importance (Bylander 2002), random forest algorithm has been frequently used in bioinformatics for variable selection, prediction modelling, pathway analysis, genetic association, epistasis detection, etc (Diaz-Uriarte and de Andrés 2005, Chen and Ishwaran 2012). RF is effective in “large p, small n” problems - situations when there are many more variables than observations and it has a good predictive performance even when most predictive variables are noisy. In addition, it does not overfit, is invariant to monotone transformations of the predictors and can handle a mixture of categorical and continuous predictors which interactions it can incorporate in the final model. Lastly, its “grouping

property" of trees enables RF to adeptly deal with correlation and interaction among variables (Ishwaran et al. 2010).

Distance correlation

The main limitation of the previously introduced Pearson and Spearman correlation coefficient is that they are only sensitive to linear relationships. To account for the nonlinear regulatory relationships, we implemented the distance correlation coefficient as well (Székely et al., 2007). The DC has proven its power and computational effectiveness (Gorfine et al., 2012) in detecting nonlinear dependence for two variables with arbitrary dimensions because its estimations are quite simple without any distribution assumption (Guo et al., 2014). In biology, it was frequently used to; for example, infer the gene regulatory networks from expression data.

In general, the key idea of distance correlation is to measure the discrepancy between the joint characteristic function and the product of its marginal characteristic functions in a special weighted L_2 space. Specifically, for random variables *(X,Y)*, denote the joint characteristic function of *(X,Y)* by *f(X,Y)*, and its marginal characteristic functions f_X and f_Y.

Pearson correlation

Pearson correlation coefficient is a measure of linear correlation between two variables. It has a value between +1 and −1, whereas +1 denotes a total positive linear correlation, 0 no linear correlation, and −1 a total negative linear correlation (Pearson, 1895).

Given a pair of random variables (X,Y) population Pearson correlation coefficient (ρ) is calculated as the covariance between random variable over the product of their standard devitations(σ_X is standard deviation of X, and σ_Y is the standard deviation of Y). The sample Pearson correlation coefficient (r_{xy}) is obtained by substituting estimates of the covariances and variances based on a sample into the formula above.

Spearman correlation

The Spearman correlation coefficient (ρ) is defined as the Pearson correlation coefficient between the rank variables. It is a nonparametric measure of rank correlation - a statistical dependence between the rankings of two variables that assesses how well the relationship between two variables can be described using a monotonic function. Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other and when there are no repeated data values. For a sample of size *n* raw scores X_i , Y_I are converted to ranks rg X_i , rgY_I and Spearman correlation coefficient is calculated as the covariance between ranks over the product of standard devitations of the ranks ($\sigma rg_X$is standard deviation of rg_X, and $\sigma_{rgY}$is the standard deviation of rg_Y).

1.6. Understanding disease variants with epigenomics

Any disease or disorder that is caused by mutations in one or more genes can be referred to as a genetic disease (Nyhan and Sakati 1987). For example, putative causal variants in apolipoprotein E have been found to be associated with Alzheimer's disease (Sillén et al. 2008), IL23R with Crohn's disease (Duerr et al. 2006), whereas a common human translocation involving *c-myc* is critical to the development of most cases of Burkitt lymphoma (Finver et al. 1988).

Since the first disease-associated enhancer was identified in Hirschsprung disease (Grice et al. 2005), more and more SNPs have been annotated to their putative causal genes based on the overlap with regulatory regions of the annotated gene (Styrkarsdottir et al. 2018; Short et al. 2018; Zhang et al. 2018; Schork et al. 2019). Soon, it became obvious that identification of regulatory elements and linking them to genes, pathways, and cellular processes represents the fundamental limitation of the functional characterization of GWAS results (and other association studies; Gallagher and Chen-Plotkin 2018), and a major obstacle for improving our understanding of diseases etiologies (Whalen et al. 2015).

To study inherited genetic variation in health and disease, traditional genetic approaches such as linkage analysis (Lathrop et al. 1984) and genome-wide association studies have been utilized (Haines et al. 2005). However, development of the whole-genome and whole-exome sequencing enabled researchers to study the role of *de novo* mutations in an unbiased manner (Veltman and Brunner 2012). Nonetheless, the great majority of identified SNPs was found to be non-coding and the precise molecular mechanisms by which those polymorphisms exert their effects remains mostly unknown (Pickrell 2014). Putative causal genes were commonly annotated to SNPs based on their proximity to genes (Welter et al. 2014) and/or overlap with eQTLs (GTEx Consortium et al. 2017). However, since the role of enhancers in genetic susceptibility to various human traits and diseases became more evident (Smith and Shilatifard 2014; Chen et al. 2018), SNPs started to be more frequently annotated to their putative causal genes based on the overlap with regulatory region of an annotated gene (Styrkarsdottir et al. 2018; Short et al. 2018; Zhang et al. 2018; Schork et al. 2019).

In my thesis, I focus on the annotation analysis of the results of genome-wide association studies (GWASs), because, during more than a decade-long history of GWAS experiments (Haines et al.

2005), thousands of SNPs have been associated with various phenotypes and diseases (Welter et al. 2014).

1.6.1. Basics of genome-wide association study (GWAS)

Genome-wide association study (GWAS) is an observational type of a study that is performed to identify genetic association with phenotypes of interest. GWASs typically focus on associations between single-nucleotide polymorphisms (SNPs) and phenotypes like human diseases. In GWASs, statistical associations are tested across many individuals in the genome-wide manner. GWASs mainly identify risk-associated "common variants" - single-nucleotide polymorphisms (SNPs) relatively frequent in the human population. In other words, GWA studies follow the "common disease-common variant" (CD/CV) hypothesis that assumes that common diseases are caused by common variants, which manifest frequently in the studied population, and lead to susceptibility to complex polygenic diseases (Schork et al. 2009).

GWAS SNPs were found to be statistically over-overrepresented in (human) diseases and traits (Freedman et al. 2011; Blattler et al. 2014), but the majority (~93%) of risk-associated index SNPs (and SNPs that are in high LD to index SNPs) are not precisely located in the protein coding regions (Tak and Farnham 2015; ENCODE Project Consortium 2004; Maurano et al. 2012). Nevertheless, noncoding variants were shown to cause common diseases more likely than the non-synonymous coding variants (Manolio et al. 2008), and they account for the majority of disease heritability (Frazer et al. 2009).

It was hypothesized that noncoding GWAS SNPs modulate disease etiology by causing changes in gene expression (Gerasimova et al. 2013). Recently, information produced through the biochemical surveys of the human genome (e.g., ENCODE (ENCODE Project Consortium 2004), Roadmap Epigenomics (Roadmap Epigenomics Consortium et al. 2015), the IHEC (Stunnenberg et al. 2016) raised high hopes that many non-coding risk-associated SNPs could be explained through their effect on gene expression. GWAS SNPs were found to be equally proportioned between the intergenic and intronic compartments (Freedman et al. 2011; Blattler et al. 2014), enriched for enhancer-associated regulatory chromatin states in biologically-relevant cell types (Ernst et al. 2011), evolutionarily conserved elements (Lindblad-Toh et al. 2011), histone marks (Schaub et al.

2012; Trynka et al. 2013; Grubert et al. 2015) and accessible regions (Maurano et al. 2012; Pickrell 2014).

Several methods that integrate epigenetic and genetic information have been developed: RegulomeDB (Boyle et al. 2012), HaploReg (Ward and Kellis 2012), FunciSNP (Coetzee et al. 2012), GWAS3D (Li et al. 2013), rSNPBase (Guo et al. 2014), etc. In addition multiple sources of enhancer-gene associations have been made available: EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018), HANCER (Wang et al. 2019), etc. However, we do not know if the initial choice of the annotation tools/data sources have a large effect on the interpretation of the results of a downstream GWAS analysis.

1.6.2. The GWAS Catalog and DisGeNET

Since the research on the genetic causes of disease has accelerated, as a result of both the completion of the human genome (Venter et al. 2001; McPherson et al. 2001; Collins et al. 2003) and the development of the NGS techniques (Ronaghi et al. 1996; Margulies et al. 2005), systematic integration of fragmented and heterogeneous information about the genetic causes of disease - scattered across specialized catalogues focused on specific disease classifications (i.e. Mendelian, or rare diseases), different model organisms, or on particular technological approaches (such as GWAS) - was crucial to support the development of precision medicine and drug discovery (Piñero et al. 2017).

One of the largest and comprehensive collections of human gene-disease associations (GDAs) currently available is DisGeNET (Piñero et al. 2017). Contrary to the GWAS Catalog, which focuses only on common polygenic diseases and variants, DisGeNET systematically integrates data from animal models, manually curated repositories, and the scientific literature. For example, The DisGeNET release 4.0 includes the following resources: the Comparative Toxicogenomics Database (CTD), UniProt, OMIM, ClinVar, Orphanet, The GWAS Catalog, the Rat Genome Database (RGD), the Mouse Genome Database (MGD), and the Genetic Association Database (GAD).

On the other hand, the GWAS Catalog is a free online database that compiles data of GWASs and summarizes their unstructured data (and data from different literature sources) into easily accessible data (Welter et al. 2014). For example, the catalog contains information and locations

of risk-associated SNPs supplemented with information about mapped/reported genes, details of the GWAS analysis (cohort size, ethnicity, etc.), study groups design, publication history, etc. After the first GWA study was performed in 2005 (Haines et al. 2005), thousands of GWASs were published and several thousands of genomic polymorphisms have been statistically associated with human phenotypes and diseases in the meantime (Hindorff et al. 2009). To organize large amounts of information produced by the GWA studies, attempts have been made at creating comprehensive catalogues of SNPs (Hindorff et al. 2009; Welter et al. 2014).

1.7. Contribution of this thesis

Despite the tremendous progress in understanding how enhancers tune gene expression and of which genes, the field still lacks an approach that is systematic, integrative and accessible for discovering and documenting *cis*-regulatory relationships across the genome. Up today, multiple approaches have been utilized to study enhancer-mediated long range gene regulation and they can be broadly categorized into four categories: predictions using information from the eQTL studies (Rockman and Kruglyak 2006; Gaffney et al. 2012; GTEx Consortium et al. 2017) or 3C-related technologies (Dekker et al. 2002; Simonis et al. 2006; Dostie et al. 2006; Lieberman-Aiden et al. 2009; Fullwood et al. 2009), and direct functional confirmation of enhancer activity by reporter assays or cellular screens (Arnold et al. 2013; Kwasnieski et al. 2012; Arnold et al. 2013; Kheradpour et al. 2013; Kvon 2015; Gasperini et al. 2019).

The fourth approach, computational modelling of *gene expression ~ enhancer activity*, was the main subject of this thesis. Recently, several data integration approaches aimed to predict enhancer-gene associations were developed based on a large number of tissues, cell types and cell lines: EnhancerAtlas (Gao et al., 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al., 2017), FOCS (Hait et al., 2018), HANCER (Wang et al. 2019). However, their results, e.g. sets of EGA predictions, were never systematically analyzed, nor were their differences characterized. We do not know whether they predict the same regions in the genome as enhancers. Likewise, we do not know if they predict the same set of enhancer-gene associations. Lastly, although enhancer-gene associations have been used to assign non-coding risk SNPs to their target genes, it is still unknown if (and how) various sets of predictions (EGAs) influence the result of risk-SNP functional analysis.

I analyzed four main problems of enhancer-gene associations: first, I reviewed and characterized differences between published sets of enhancer-gene associations, and then, I set off to develop a novel approach that models and integrates *gene expression ~ enhancer activity* (reg2gene). I performed a thorough benchmarking of seven sets of EGAs, and lastly, I reviewed how different predictions of enhancer-gene association could influence the results of functional analysis of SNPs. Thus, to get a complete overview of the complex problem such as associating genes with their regulatory regions, I supplemented the original idea - development of a novel tool to link

enhancers with genes they regulate - with many additional information and somewhat unexpected analyses.

In **Chapter 2**, I describe datasets and methods that I used/integrated to map and benchmark enhancer-gene associations.

In **Chapter 3**, I introduce the problem of computational modelling of *gene expression ~ enhancer activity.* I review four computational methods that predict EGAs: JEME - joint effect of multiple enhancers (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS - FDR-corrected OLS with Cross-validation and Shrinkage (Hait et al. 2018). I showed that individual sets of predictions differ tremendously; especially in the location, number and properties of defined enhancer regions and characteristics of enhancer-gene associations.

In **Chapter 4**, I describe our work to develop a computational method that models and integrates *enhancer activity ~ gene expression* is the genome-wide and systematic manner - reg2gene. Although our central question was to computationally model enhancer-gene association, due to a prominent role of enhancer definition and its influence on the final modelling performance, I dedicated a large section of this chapter towards reviewing, analyzing and identifying the robust set of enhancers. I thoroughly explain the process of developing the reg2gene method and characterize three sets of EGAs (enhancer-gene associations) that were identified using reg2gene data modeling and integration: *stringentC, flexibleC, inhouseM.*

The work described in **Chapter 5** provided us with additional information about seven sets of (analyzed or developed) enhancer-gene associations: EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), and FOCS (Hait et al. 2018) and three *inhouse* models: *stringentC, flexibleC, inhouseM.* Specifically, I ran benchmarking of the reported associations by overlapping them with commonly used benchmark datasets: eQTLs (Westra et al. 2013, GTEx Consortium et al. 2017) and chromatin interactions reported in Xie et al. (2016) and Javierre et al. (2016). I further benchmarked EGAs based on results of cellular screens (Gasperini et al. 2019) and an *inhouse* method that identified false positives (FP) and true negatives (TN) in the genome. I show that the benchmark datasets suffer from low-reproducibility rate and that the *stringentC* models have the highest PPV (positive predictive values) if cellular screen and inhouse set of negative EGAs were used to benchmark data.

In **Chapter 6**, I present results of SNP-to-gene annotation analysis performed using different sources of enhancer-gene associations. Specifically, I annotated risk polymorphisms from the GWAS Catalog, colorectal cancer (CRC) SNPs, rs104111210 and showed that sets of annotated genes varied in their size. I demonstrate that some well-known gene-CRC associations were identified by certain sets of EGAs, while they were found to be missed by other EGAs methods. Lastly, I detected examples of enhancer-based pleiotropy and emphasized the potential of EGAs to identify it. I reported possible TFs that underlie such regulation and identify a novel gene-disease association using overlap of SNPs with enhancers.

In **Chapter 7**, I summarize the main results of this thesis and conclude with my view on the remaining perspectives of computational modelling of enhancer-gene associations. I elaborate the outlook for applying our growing understanding of transcriptional regulation to dissect the contributions of noncoding genetic variation to human disease and to manipulate gene expression for therapeutic purposes.

2

Methods

I performed all computational modelling and downstream analysis supervised by Dr. Vedran Franke and Dr. Altuna Akalin. They contributed with their comments, suggestions and discussions.

2.1. Data processing and integration

As a computational and data integration tool, reg2gene, relies on extensive data integration of multiple (epi)genetic datasets. Herby, I shortly introduce those datasets.

The NIH Epigenomic Roadmap download and preprocessing

We used epigenomes reported in the NIH Epigenomic Roadmap dataset - at time the largest collection of human epigenomes for primary cells and tissues - to define enhancers and predict enhancer-gene associations (Bernstein et al., 2010; Roadmap Epigenomics Consortium et al., 2015). Its 111 reference human epigenomes were profiled for histone modification patterns, DNA accessibility, DNA methylation and RNA expression; and were additionally extended with sixteen epigenomes from the ENCODE Project (ENCODE Project Consortium, 2004). Specifically, we selected epigenomic datasets - H3K27ac, H3K4me1, DNase I hypersensitivity, and DNA methylation. In addition, we used ChromHMM-predicted chromatin states for 127 Roadmap epigenomes (15-state ChromHMM model; (Ernst and Kellis, 2012). ChromHMM aggregated multi-dimensional matrices of chromatin marks into a small number of chromatin states based on a multivariate Hidden Markov Model and was previously used to systematically characterize chromatin states across 127 Roadmap cell types.

From the https://egg2.wustl.edu/roadmap/, we downloaded 127 imputed H3K27ac and H3K4me1 ChIP-Seq, DNase-Seq, RNA-Seq, and WGBS DNA fractional methylation genome-wide signal coverage tracks from the NIH Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015), as bigWig files. We removed epigenomes reported to be of poor quality (three WGBS fractional methylation reference epigenomes: E001, E003, and E017). We allowed for only one donor per epigenome tissue or cell type. In addition, we downloaded info about spatial context (chromatin states) across 127 epigenomes from the core 15-state model, that was prediceded using ChromHMM v1.10 (Ernst and Kellis, 2012) https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz.

Computational models of enhancer-gene associations: JEME, FOCS, GeneHancer, EnhancerAtlas

GeneHancer preprocessing

From https://genecards.weizmann.ac.il/geneloc/index.shtml, we downloaded single .csv document (GeneHancer enhancers are not cell-type specific) with 243,281 GeneHancer enhancers on the 28.11.17. GeneHancer version was 4.6 and it contained enhancer locations and corresponding metadata; including names of associated genes. We wrote an inhouse R scripts to: 1) extract locations of enhancers, 2) obtain one-to-one mapping between genes and enhancers (necessary since more than one gene was reported to be associated with single enhancer region saved as one row of the table), 3) liftover hg38 to hg19 enhancer coordinates (rtracklayer R package), 4) obtain gene locations either by using GeneHancer-reported HCGN gene names or ENSEMBL IDs and setting them to be equal to the extended TSS coordinates (+/-1000 bp) of the GENCODE v24 genes. Post processing, we defined 506,684 unique GeneHancer unique enhancer-promoter associations.

EnhancerAtlas preprocessing

We downloaded enhancer-target relationships for 70 EnhancerAtlas cell/tissue types from http://www.enhanceratlas.org/ on the 28.11.17. We used an in-house script to create a union of all enhancer-target relationships reported in different cell types, by pooling all reported interactions across cell types into one object. To account for the multiple observations of the same E - G association across multiple cell/tissue types, we requested that each enhancer-target relationship can occur only once in the final dataset. Since only ENSEMBL transcript identifiers were reported, we performed ENST mapping to ENSG using EnsDb.Hsapiens.v75 to obtain gene IDs. Based on the overlap in ENSEMBL ID's between GENCODE and reported genes, EnhancerAtlas gene coordinates were set to be equal to extended TSS coordinates (+/-1000 bp) of the GENCODE v24 genes.

FOCS preprocessing

We accessed FOCS enhancers (on the 15.1.2018.) and enhancer-promoter links (on the 13.9.2018) from the http://acgt.cs.tau.ac.il/focs/download.html. We downloaded all four

reported datasets of enhancer predictions: 1) ENCODE DHS (encode.enh.pos.RData), 2) Roadmap DHS (roadmap.enh.pos.RData), 3) FANTOM5 CAGE (fantom.enh.pos.RData), and 4) GRO-Seq enhancer regions (groseq.enh.pos.RData).

Additionally, we downloaded four reported sets of enhancer - gene FOCS associations: 1) ENCODE DHS (encode_interactions.txt), 2) the Roadmap Epigenomics project DHS (roadmap_interactions.txt), 3) FANTOM5 CAGE (fantom_interactions.txt), and 4) GRO-Seq (groseq_interactions.txt), and pooled reported interactions together. We wrote an in-house script to assign coordinates to the FOCS gene names by overlapping reported gene names with GENCODE v24 genes and taking over GENCODE extended (+/-1000 bp) TSS coordinates. After excluding interactions on the chromosome Y, we obtained 117,355 unique FOCS EP links.

JEME preprocessing

We downloaded four JEME datasets on the 11 October 2017 from http://yiplab.cse.cuhk.edu.hk/jeme/: elastic net and LASSO predictions (enhancer - gene links) for FANTOM5 and ENCODE/Roadmap cell types and tissues. To define JEME enhancers, we pooled results of LASSO and elastic net predictions across all Roadmap cell types and tissues, and extracted reported enhancer regions and kept unique entries. In addition, for analysis of enhancer-gene interactions, we pooled all predictions across all four JEME datasets and identified 929,682 unique enhancer-gene associations. Gene coordinates were set to be equal to the extended TSS coordinates (+/-1000 bp) for genes overlapping the GENCODE v24 genes.

Benchmark datasets: eQTL and studies of chromatin interactions

PC-HiC dataset preprocessing

(Javierre et al., 2016) reported a genome-wide dataset that is strongly enriched for genomic interactions with promoter regions across 17 human primary blood cell types. It was experimentally created by PC-HiC methods. Active promoter enhancer links from PC-HiC dataset were accessed as a PCHiC_peak_matrix_cutoff5.tsv object on 17.1.2017. from http://www.sciencedirect.com/science/article/pii/S0092867416313228. We pooled interactions across cell types and kept only unique interactions. We identified a total of 728,838 unique interactions (31,253 annotated promoters and on average 175,000 interactions per cell type).

CCSI preprocessing

We accessed the CCSI (Chromatin Chromatin Space Interaction) database on the 11th January 2017 from http://120.79.23.67/ccsi/download.php and downloaded all hg38 4C, 5C, HiC, ChIA-PET datasets (N=44). We pooled interactions from all reported datasets into one object, and lifteovered (to the human genome version hg19) coordinates of both anchors of all interacting pairs using CrossMap program. We removed all interacting pairs: 1) which had the start and end coordinate of the anchor switched (start was downstream of end); 2) at least one anchor of the interacting pair was longer than 10 kb, or 3) two anchors within the same interacting pair overlapped together. In the original publication, a total of 3 017 962 pairwise interactions across 91 chromatin interaction datasets was reported. Since eQTLs occupy a single position in the genome, whereas anchors of interactions can vary in size, we tested the size of interacting regions reported in CCSI database and PC-HiC interactions. We identified large differences in the size of reported interacting regions (**Supplementary Figure 1.**). The CCSI database reported interactions among 1Mb long regions in the genome. Since long regions more likely overlap one or more enhancers (that themselves generally span several hundred base pairs), we decided to set a limit on the size of interacting regions - 10Kb and identified 1 587 002 such interaction pairs.

TAD regions

We downloaded coordinates of TADs regions for 37 hg19 cell types/tissues on the 25th of April 2019 from the **http://promoter.bx.psu.edu/hi-c/publications.html**. TADs coordinates originated from five publications (Lieberman-Aiden et al. 2009, Rao et al. 2014, Dixon et al. 2015, Leung et al. 2015, Schmitt et al. 2016) and as a part of ENCODE Consortium (ENCODE Project Consortium 2012).

GTEx database preprocessing

We accessed 44 GTEx datasets V6 (corresponding to 44 analyzed cell types) on 5.10.2016. from the http://gtexportal.org/home/datasets. We pooled together info about eQTLs and associated gene names (.v6p.signif_snpgene_pairs.txt) across all 44 tissues into one object, and separately pooled info about genes (.v6p.egenes.txt) into another object. We merged these two datasets together by gene's id and identified a total of 6,654,931 entries, or 2,498,498 unique eQTL-gene pairs. Gene coordinates were set to be equal to extended TSS coordinates (+/- 1000bp) identified in the GENCODE dataset for genes which ENSEMBL IDs overlapped. We pooled together EGAs across all cell types and identified 1 million *cis* eQTLs-gene associations (N=1 034 370) of which 388,160 was unique (**Supplementary Figure 2.**).

Westra et al. (2013) eQTLs preprocessing

We downloaded trans- and cis- eQTLs reported in Westra et al., (2013) on 9.3.2017. from www.genenetwork.nl/bloodeqtlbrowser/ (2012-12-21-CisAssociationsProbeLevelFDR0.5, 2012-12-21-TransAssociationsProbeLevelFDR0.5). We liftovered hg18 to hg19 eQTL coordinates using rtracklayer R package and we set gene coordinates to be equal to extended TSS coordinates (+/-1000 bp) of the GENCODE v24 genes based on the overlap in ENSEMBL ID's between GENCODE and reported genes. In Westra et al., (2013), 1,962,237 *cis* and 4,542 *trans* eQTLs from peripheral blood samples of several thousand individuals was reported however, after processing 672,717 eQTLs remainded.

"CRISPR/Cas9 positives" (Fulco et al., 2019; Gasperini et al., 2019) preprocessing

We downloaded table of the 664 enhancer-gene pairs on the 15 January 2019 (https://www.cell.com/cms/10.1016/j.cell.2018.11.029/attachment/d650f1e9-c627-4073-a6db-6fe44d4e2149/mmc2.xlsx). We used an in-house script to pair enhancers and genes that reported in separate tabs of the same object. Since gene coordinates were not reported, based on reported ENSEMBL ID, we identified and set genes coordinates to be equal to extended TSS coordinates (+/- 1000bp) reported for genes in the GENCODE dataset. This resulted in 449 "high-confidence" enhancer-gene interactions.

GENCODE preprocessing

GTF files were downloaded from http://www.gencodegenes.org/releases/24lift37.html (01.06.2016.), and liftovered from GRCh38 to GRCh37 using in-house script. GTF files for basic annotation and non-coding RNAs were merged together. Genes located on the Y chromosome were removed from further analysis due to the lack of chrY coverage in some of the reported Roadmap epigenomes. Exon locations (GTF entries with type "exon") with confidence levels 1 and 2, and exon length longer than 10bp were used in the further analysis. In addition, exon regions were reduced (for each gene we individually reduced its exons), such that, per individual gene, each location in the genome could be covered with only one exon. For each gene, we assessed the corresponding transcriptional start site, and extended it +/-1000bp (promoters function from GenomicRanges R package).

TF download and preprocessing

We downloaded human TF binding sites in narrowPeak format and corresponding metadata from the UCSC Uniform track http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/ on the 6th of September 2016. A total of 690 narrow peak profiles corresponded to 161 TFs.

GWAS Catalog preprocessing

We accessed and downloaded GWAS Catalog on the 5th of February 2019, using makeCurrentGwascat() function from the gwascat R package. We liftovered SNP coordinates from the hg38 to hg19 using liftOver() function from rtracklayer. Since we were particularly interested in the non-coding polymorphisms we screened the "CONTEXT" reported in the GWAS Catalog and extracted only SNPs in one of the following categories: intergenic variant, intron variant, 3 prime UTR variant, 5 prime UTR variant, regulatory region variant, upstream gene variant, TF binding site variant, TF binding site variant x intron variant. In addition, we removed all entries that overlapped exons of protein coding genes (exon regions from GENCODE, used in the previous analyses).

DISGENET download and preprocessing

We downloaded all variant-disease associations from the DIGENET database on 25.7.2019. from http://www.disgenet.org, and filtered out associations reported in the GWAS Catalog or GWASDB. This resulted in 310,502 DISGENET entries. To cross-reference phenotype terms from DISGENET and GWAS Catalog we downloaded the Experimental Factor Ontology (EFO) file in the OBO format (https://www.ebi.ac.uk/efo/efo.obo; 25.7.2019.).

2.2. Characterization of published enhancers and enhancer-gene associations

2.2.1. Extracting and characterizing enhancer regions

Locations of enhancer regions were individually extracted from the processed sets of enhancer-gene associations for each analyzed method: EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), and FOCS (Hait et al. 2018). Unique enhancer regions were used for downstream comparisons. In the case of FOCS enhancers, we separately extracted four different enhancer definitions: GRO-Seq, ENCODE, Roadmap, FANTOM5. Nevertheless, when compared to other sets of enhancers, we used only enhancer regions that were reported in the

final FOCS models. For each enhancer definition, we quantified their number, size distribution, genome coverage, etc.

2.2.2. Identifying an overlap between enhancers and other functional elements

We assessed overlaps between enhancer regions and exons, introns, intergenic regions and promoters in the human genome.

First, for each gene, we identified exon locations in the GENCODE database (Harrow et al., 2012) and infer other genomic features. We defined promoters as regions 2000bp upstream and 200bp downstream from the TSS of a gene. All gaps between exons but within genes were assessed as introns, whereas intergenic regions were assessed as gaps between genes. We quantified the number of overlaps between enhancers and all four categories of genomic features.

2.2.3. Juxtaposing sets of enhancers

To assess an overlap between four different sets of enhancers we first calculated the **Jaccard index** (as the pair-wise measure of similarity between sets) in an inhouse script that, for each combination of enhancers, identified an overlap between enhancers (binary output), and divided the size of their intersection by the size of their union. An overlap between enhancers was assessed using the GenomicRanges package in R.

In addition, we created a union of all enhancer definitions together and identified all base pairs in the genome that are "covered" by at least one enhancer definition. Next, we quantified the enhancer coverage for those positions. In other words, we identified groups of overlapping enhancer elements and reduced them to one enhancer region whose start and end positions were based on the lowest start and highest end positions within its group of overlapping enhancer elements. For each unified enhancer, we identified publications which reported enhancer(s) that overlap(s) with a unified enhancer definition. Each unified enhancer definition can have an overlapping enhancer in one (publication-specific enhancers) or up to four publications.

2.2.4. Analyzing enhancer activity signals for sets of enhancers

To analyze differences in enhancer activity for different enhancer definitions we quantified their activity based on the signal from the genome-wide H3K27ac, H3K4me1 ChIP-Seq and DNase-Seq

and WGBS coverage tracks from the Roadmap dataset (Bernstein et al., 2010). Individually for each dataset, we extracted enhancer regions on the chromosome 1 and calculated the mean coverage value of the input signal in each Roadmap cell types (for example a ChIP-Seq signal for certain histone modification) using regActivity() function from the reg2gene R package. For each cell type and chromatin mark, we reported a single number - an averaged quantified enhancer signal for chromosome 1.

2.2.5. Comparing predicted enhancers and ChromHMM-predicted chromatin states

For each enhancer definition, we used enhancers on the chromosome 1 as windows and screened the ChromHMM-predicted (Ernst and Kellis, 2012) 15-state chromatin models across 127 Roadmap epigenomes (Ernst and Kellis, 2012). For each identified chromatin state, we quantified the percentage of per-base coverage with enhancers. We did that individually for each cell type and then averaged the signal across cell types. Later, the same analysis was repeated for the "in-house" and "consensus" enhancers.

2.2.6. Identifying an overlap between sets of enhancer-gene associations

To identify intersections between two sets of enhancer - gene associations reported in different publications, we used benchmarkInteractions() function from the reg2gene package. This function compares two datasets at the time, and identifies if locations of both elements of the pair, enhancers and genes, from those two analyzed datasets overlap together. If they do, they were considered to be overlapping counterparts. We requested that the result of this analysis is binary. Importantly, two enhancer regions (from the two tested enhancer-gene prediction methods) had to have an overlap in at least one 1bp, whereas the overlap of the two associated genes was measured on the level of TSS and TSSes were requested to be equal. To adjust for the fact that gene locations were not uniform across different publications (for example, JEME gene locations were reported as TSSes, whereas we identified the GeneHancer gene locations using reported names and the biomaRt functions; Durinck et al., 2009), we equalized gene locations based on their ENSEMBL IDs. Specifically, we used gene names reported in each publication (ENSEMBL IDs), matched them with the GENCODE genes, and used the matched GENCODE TSS locations as enhancer-associated gene locations. We resized TSSes +/- 1000 bp.

For each set of enhancer-gene associations, we identified an overlap with all other datasets, and summed up the number of identified overlaps across all five publications. Since we requested that

overlap information is binary (whether or not there is at least one overlapping E-G association), for a single enhancer - gene association we could report from 1 to 4 counts. For example if one enhancer - gene association from FOCS, has at least one counterpart in JEME and EnhancerAtlas, it will receive a count of three (or EGA overlapped with FOCS, EnhancerAtlas and JEME). Later, the same procedure was repeated for the "in-house" and "consensus" EGAs.

2.3. reg2gene

2.3.1. reg2gene algorithm

reg2gene method is mostly explained in Chapter 4.

2.3.2. reg2gene R package

reg2gene is available as an R package at the Github: https://github.com/BIMSBbioinfo/reg2gene

2.3.3. Defining "in-house" enhancers

We defined 184,005 "in-house" enhancers based on the cross-cell-type information from the core 15-state ChromHMM model v1.10 (Ernst and Kellis, 2012). Briefly, ChromHMM aggregated multi-dimensional matrices of chromatin marks into a small number of chromatin states and systematically characterized those states across 127 Roadmap cell types. Specifically, we calculated per base coverage of ChromHMM EnhG_6 and Enh_7 chromatin states across 127 Roadmap epigenomes (Kundaje et al., 2015). Stretches of the genome predicted as EnhG_6 and Enh_7 in at least 20% of the Roadmap epigenomes, e.g. minimum per base coverage of 24, were identified as enhancers.

2.3.4. Unifying enhancer definitions into "consensus" enhancers

First, a "consensus" enhancer definition was obtained by calculating an overlap between our own ChromHMM-based "in-house" definition and previously proposed enhancer definitions from EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), and FOCS (Hait et al. 2018). Importantly, we did not use one enhancer definition from FOCS whereas we separately included: four groups of FOCS enhancers: DHS Roadmap, DHS ENCODE, GRO-seq,

and CAGE FOCS enhancers. In addition, we subsetted enhancers from EnhancerAtlas and included only those that were smaller than 5kb and JEME enhancers reported in the elastic net models. As explained in Chapter 2.1., within each publication, we merged/reduced together all overlapping enhancers, e.g. groups of enhancers that overlapped together were defined as one enhancer region which start and end positions were based on the lowest start and highest end positions of original enhancers).

As a result, we calculated the per base genome coverage using eight individual enhancer definitions ("in-house", JEME, GeneHancer, EnhancerAtlas and four FOCS enhancer definitions: DHS Roadmap, DHS ENCODE, GRO-seq, and CAGE FOCS). We defined "consensus" enhancers as stretches of the genome covered by at least three enhancer definitions.

Second, we additionally processed "consensus" enhancers as follows: first, we overlapped enhancers with the binding sites of 161 transcription factors (TFBS) designated by ENCODE ChIP-seq peaks (POLR2A and POLR3G were excluded). If enhancers overlapped TFBS, and boundaries of corresponding ChIP-seq peaks were longer than the enhancer region boundaries, we extended enhancer boundaries to match those of TFBS. We additionally extended enhancers for +/- 100bp, merged them, and then reduced them back to their original width.

We removed enhancer regions smaller than 20bp, and processed large (> 2kb) enhancer regions in the three steps procedure in which we used borders between high and low TF occupancy in the genome to segment large enhancers into smaller functional subunits. TF occupancy borders were identified based on the processed ENCODE ChIP-seq peak tracks.

We calculated genome-wide per base TF coverage for 161 transcription factors (ENCODE ChIP-seq narrow peaks excluding POLR2A and POLR3G). We used the calculated median of base-pair coverage as a cutoff to segment coverage track into two types of regions - regions with high and low TF occupancy (median of 23 was calculated exclusively using large enhancers). We partitioned large enhancers into smaller functional subunits (<2kb) using inferred borders between high and low TF occupancy and functional segments with above median ChIP-seq peak coverage were retained for further analysis, whereas those with below median ChIP-seq peak coverage were removed.

Small enhancer regions (<50bp), created during the segmentation step, were merged with the first upstream enhancer region (alternatively to the first downstream enhancer). Enhancers were extended +/-100bp, merged, and reduced to the previous size.

Large enhancers that did not overlap with borders of high and low TF occupancy were tiled to 2kb segments. Smaller identified functional segments with above median ChIP-seq peak coverage were retained for further analysis.

In the end, we re-checked distribution of the width of enhancers, and all short enhancers were removed. In addition, we removed all enhancers located 2 kb upstream and 1kb downstream of the genome-wide locations of TSSes when such enhancers did not overlap the first intron. We further characterized enhancers as described in Section 2.2.5.

2.3.5. Testing gene expression quantification protocol

To test a precision of the reg2gene signal quantification protocol, we compared the previously reported scores and reg2gene calculated values. Specifically, we focused on the gene expression scores calculated based on the RNA-Seq signals. We downloaded RPKMs calculated by the Roadmap consortium for all 127 cell types and compared them with levels of gene expression quantified using the bwToGeneExp function from the reg2gene R package. For each cell type individually, we quantified gene expression and calculated the Pearson correlation coefficient between identified gene expression levels and reported RPKMs. We reported correlation statistics across all cell types.

2.3.6. Analyzing an overlap with TADs

We extracted coordinates of TADs regions for 37 hg19 cell types/tissues and originated from five publications (Lieberman-Aiden et al., 2009, Rao et al., 2014; Dixon et al., 2015; Leung et al., 2015; Schmitt et al., 2016) and as a part of ENCODE Consortium (ENCODE Project Consortium, 2012). We identified whether a given enhancer-gene association (EGA) is located within any of the defined TAD regions. For each cell type, we calculated the percentage of EGAs that were found within TAD regions, and determined distribution across cell-types (presented as a histogram).

2.4. “Benchmarking” enhancer-gene associations

2.4.1. Analyzing an overlap with eQTLs

In detail explained in Chapter 5.

2.4.2. Analyzing an overlap with chromatin interactions

In detail explained in Chapter 5.

2.5. Benchmarking with “positive” and defining “negative” EGAs

2.5.1. Defining “negative” EGAs

In detail explained in Chapter 5.

2.5.2. Defining “positive” EGAs

In detail explained in Chapter 5.

2.6. Functional analysis of GWAS Catalog SNPs, CRC SNPs and rs10411210

2.6.1. Annotating the GWAS Catalog

We annotated each GWAS Catalog polymorphism to the target gene based on the hierarchical overlapping procedure. First, genomic regions of interest were annotated as promoters and associated with corresponding genes if they were located within +/-1000bp from the TSS. Such SNPs were filtered out. The remaining genomic regions were overlapped with enhancer regions, and genes associated with those enhancer regions were assigned to polymorphism.

2.6.2. Identifying the nearest gene for each SNP in the GWAS Catalog

Using the GENCODE-reported TSS, we wrote an inhouse R script that calculates the distance between TSS and SNP, and assigns closest gene to the analyzed SNP.

2.6.3. Identifying an overlap between enhancers and other functional elements

We identified an overlap between SNPs and other functional elements in the human genome as described in Section 2.2.2.

2.6.4. Visualizing results

To visualize results of SNP-to-gene annotation analysis with different sets of enhancer-gene associations (EGAs), we developed a specific R function - plotInteractions().

2.6.5. Identifying SNPs in LD with the index SNP

To identify all SNPs that are in linkage disequilibrium (LD) with the index SNP we used the proxysnps R package (https://rdrr.io/github/slowkow/proxysnps/) and get_proxies() function. First, using the SNP identifier we queried the Biomart database (grch37.ensembl.org) using the biomaRt R package (Smedley et al., 2009) and identified the genome location of a tested SNP. Then, we searched for all SNPs that are in LD with the index SNP in the CEU population and R.squared=0.6 using get_proxies() .

2.6.6. Performing enrichment analysis

For each identified set of CRC-associated genes we performed gene enrichment analysis for: molecular function, cellular component, biological process, human phenotype ontology, OMIM disease, and KEGG, Reactome, and PANTHER pathways using enrichR R package (Kuleshov et al., 2016) and identified TOP10 enrichment results.

2.6.7. Identifying transcription factor binding sites

To identify transcription factor binding sites, we wrote an inhouse function that queries genome (enhancers) sequences and identifies motifs for transcription factors that are available in the JASPAR2014 database (Mathelier et al., 2014). Our function first identifies genomic sequences for

enhancer regions using getSeq() from BSgenome.Hsapiens.UCSC.hg19 R packages, and then searches for TF binding motifs using searchSeq() functions from TFBSTools R package (Tan and Lenhard, 2016), min.score="90%", strand="*").

2.6.8. Identifying TF motifs

In addition, we wrote an inhouse function that searches through the ChIP-Seq UCSC datasets and identifies an overlap between TF binding peaks and genomic regions of interest (enhancers). A total of 690 narrow peak profiles was analyzed (it corresponded to 161 TFs). Data processing and download was explained in Section 2.1.

2.6.9. Benchmarking with the DisGeNET sets of genes

To identify colorectal cancer associated genes we queried the DisGeNET database for the EFO:0005842 term. In addition, we identified ancestral terms for colorectal cancer: intestinal cancer, carcinoma, cancer, neoplasm and queried the DisGeNET database for those terms. For that we used the ontoCAT R package (Adamusiak et al., 2011) and getTermSynonyms() and getTermById() functions.

2.6.10. Literature search

To identify gene-disease associations, I additionally search the text-mining tool DISEASE (http://diseases.jensenlab.org/).

3

Results I: Review of the computational methods that assess enhancer-gene associations (EGAs)

Preface

In this section, I review computational genome-wide methods that were developed to map cis-regulatory interactions in the human genome. Based on conversations with Dr. Altuna Akalin and Dr. Vedran Franke, I defined the scope of this analysis.

Abstract

Multiple computational methods that model enhancer activity ~ gene expression have been developed to map cis-regulatory interactions in the human genome. However, different sets of EGA predictions were never systematically analyzed, nor were their differences characterized.

Here, we analyzed four methods that computationally model EGAs: JEME - joint effect of multiple enhancers (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS - FDR-corrected OLS with Cross-validation and Shrinkage (Hait et al. 2018). We showed that individual sets of predictions differ tremendously; first in the location, number and properties of defined enhancer regions, and then, we demonstrated that methods generally predict discern sets of enhancer-gene associations. This analysis allowed us, for the first time, to systematically compare computational predictions of enhancer-gene associations and identify their differences.

We postulated that the observed changes are likely a consequence of the fact that different authors selected mostly distinct sources of data for modelling, validation, benchmarking, or to define enhancers.

3.1. Introduction

Thousands of regulatory DNA sequences (especially enhancers) have been identified in the human genome, however, elucidating their function and target genes remains elusive (ENCODE Project Consortium 2004; ENCODE Project Consortium 2012). The field lacks an approach that is systematic, integrative and accessible for discovering and documenting *cis*-regulatory relationships across the genome. Different experimental and computational techniques, as well as data integration approaches were proposed to improve our understanding of enhancer-mediated gene expression regulation (Hariprakash and Ferrari 2019).

In this research, we specifically focused on studying enhancer-mediated gene expression regulation by means of computational modelling of *enhancer activity ~ gene expression*. Multiple computational methods that study enhancer-gene associations have been developed (He et al. 2014; Corradin et al. 2014; Roy et al. 2015; Gao et al. 2016; Whalen et al. 2016; Cao et al. 2017; Fishilevich et al. 2017; O'Connor et al. 2017; Hait et al. 2018; Li et al. 2018; Libbrecht et al. 2018; Wang et al. 2019, Singh et al. 2019), but their results were never systematically reviewed. Up until today, two crucial questions remained unanswered:

a) Do computationally predicted enhancer-gene associations (EGAs) predict the same enhancer regions in the genome, or not?
b) If and how much computationally predicted enhancer-gene associations (EGAs) predictions are similar or different to each other.

To answer those questions, we analyzed four computationally models of EGAs: JEME(Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS (Hait et al. 2018).

3.2. Materials

3.2.1. An overview of used datasets - reg2gene

I reviewed several computational methods developed to predict enhancer-gene associations: JEME (Cao et al. 2017), EnhancerAtlas (Gao et al. 2016), GeneHancer (Hait et al. 2018), FOCS (Fishilevich et al. 2017), and later, in the next Chapter, I used it to define locations of enhancers and improve enhancer-gene associations (**Table 3.1.**).

Table 3.1. Characteristics and statistics behind datasets used in this chapter

Data source	Technology/Data type	Processing algorithm	Used as/for:	N[EGAs or interaction]
JEME (Cao et al. 2017)	Roadmap (H3K4me1, H3K27ac, DHS, H3K27me3) and FANTOM CAGE, ChIA-PET, Hi-C, eQTL	LASSO and elastic net, random forest	Improving enhancer-gene associations	929,682
GeneHancer (Fishilevich et al. 2017)	ENCODE (DHSs, H3K27ac), Ensembl, Roadmap, FANTOM5, VISTA, eQTLs , eRNA co-expression , TF co-expression, capture Hi-C (CHi-C) and gene target distance (nearest neighbor)	Score-based calculation of SGE	Improving enhancer-gene associations	506,471
EnhancerAtlas 2.0 (Gao et al. 2016)	ChIP-Seq, DNase-Seq, CAGE, FAIRE-Seq, transcription factor binding and DHS, FAIRE, eRNA, P300 binding sites, POL2 binding sites, and eRNA	IM-PET (distance, H3K4me1, H3K4me3, H3K27ac, gene expression)	Improving enhancer-gene associations	2.33 M
FOCS (Hait et al. 2018)	GRO-Seq, FANTOM5 CAGE, Roadmap DHSs, ENCODE DHSs	FDR-corrected OLS with Cross-validation and Shrinkage	Improving enhancer-gene associations	117,355

3.3. Results

3.3.1. Identification of computational approaches to study enhancer-gene associations (EGAs)

In order to identify previously published enhancer-gene associations and analyze their pros and cons, prior to developing the reg2gene algorithm, we set off to review all previous publications that computationally model enhancer-gene associations. We identified multiple such publications (**Table 3.2.**).

Table 3.2. Summary of computational methods used to study enhancers and enhancer-gene associations (EGAs)

Exclusion reason	Method	Author	Paper
NA	**FOCS**	Hait et al., 2018	FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map
NA	**JEME**	Cao et al. 2017	Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines
NA	**GeneHancer**	Fishilevich et al. 2017	GeneHancer: genome-wide integration of enhancers and target genes in GeneCards
NA	**EnhancerAtlas**	Gao et al., 2016	EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types.
Data integration only to define enhancers	**HACER**	Wang et al. 2019	HACER: an atlas of human active enhancers to interpret regulatory variants
12 cell lines	**IM-PET**	He et al. 2014	Global view of enhancer–promoter interactome in human cells
5 cell lines	**RIPPLE**	Roy et al. 2015	A predictive modeling approach for cell line-specific long-range regulatory interactions
12 PreSTIGE cell lines	**PreSTIGE**	Corradin et al. 2014	Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits
6 ENCODE cell lines	**TargetFinder**	Whalen et al. 2016	Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin

8 cell types	**DECRES**	Li et al. 2018	Genome-wide prediction of cis-regulatory regions using supervised deep learning methods
No enhancer-gene predictions	**dendb**	Ashoor et al. 2015	DENdb: database of integrated human enhancers
No enhancer-gene predictions	**Segway**	Libbrecht et al. 2018	A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types
8 cell lines	**CISMAPPER**	O'Connor et al. 2016	CISMAPPER: predicting regulatory interactions from transcription factor ChIP-seq data
6 cell lines	**SPEID**	Singh et al. 2016	Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks

We decided not to thoroughly review publications that had no computational component such as Segway (Libbrecht et al. 2018) and HANCER (Wang et al. 2019) and methods which models were built upon on a small number of cell lines: RIPPLE (Roy et al. 2015), TargetFinder (Whalen et al. 2016), SPEID (Singh et al. 2019), CISMAPPER (O'Connor et al. 2017), DECRES (Li et al. 2018), IM-PET (He et al. 2014) or PRESTIGE Corradin et al. 2014). For example, RIPPLE modelled five cell lines, TargetFinder and SPEID six, CISMAPPER and DECRES eight, and IM-PET or PRESTIGE modelled 12 cell lines.

In the end, we analyzed only four methods that computationally model EGAs: JEME - joint effect of multiple enhancers (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS - FDR-corrected OLS with Cross-validation and Shrinkage (Hait et al. 2018). As expected, each of them individually differed in the selection of data sets that were used to define enhancers or model enhancer-gene interactions, as well as in their choice of algorithms (**Table 3.3.**). For example, in GeneHancer, Fishilevich et al. 2017 pooled previously reported enhancers from the Ensembl database (Cunningham et al. 2015) together with FANTOM5 and VISTA (Visel et al. 2007) enhancers, whereas Hait et al. 2018 (FOCS) used DHS peak positions, raw sequence data of 245 GRO-Seq samples and CAGE tag peaks. In terms of algorithm selection, JEME, as compared to FOCS that represents the only full regression method, is a combination of regression and supervised learning methods, whereas GeneHancer reported "elite" enhancers and enhancer-gene associations based on calculated threshold score (S_{GE}) that integrated various levels of information. In addition, the number of cell types and tissues that were used for modelling differed as well across methods. Gao et al. (2016) reported associations across 48 cell

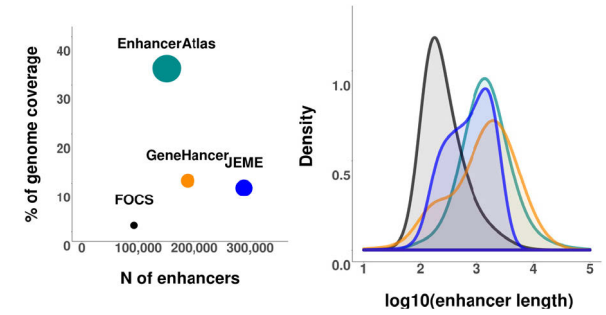
lines and 22 tissues, JEME predictions are available for two datasets: 127 Roadmap cell types and tissues and 808 FANTOM5 samples. FOCS modelled promoters and its ten closest enhancers using DHS signals from 106 ENCODE cell types, 73 Roadmap DHS cell types and tissues, 600 FANTOM5 human cell lines and primary cells, and GRO-Seq expression data from 23 different human cell lines.

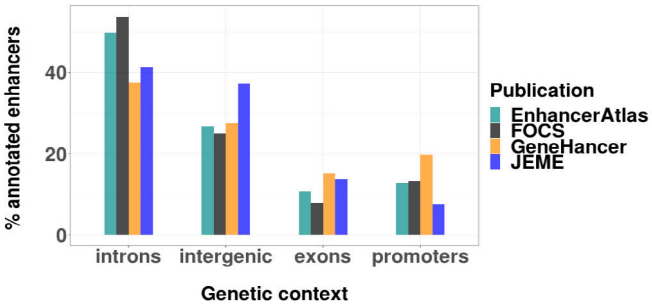
Table 3.3. Details about four computational methods analyzed in this thesis that model enhancer-gene interactions

Method	Datasets used to define enhancers	Datasets used to define EGAs	Algorithms	Benchmark datasets	N of cell types
FOCS (Hait et al. 2018)	Peaks from GRO-Seq, FANTOM5 CAGE, Roadmap DHSs, ENCODE DHSs	expression data (GRO-Seq, FANTOM5) and DHSs level (Roadmap, ENCODE)	FDR-corrected OLS with Cross-validation and Shrinkage	RNAPII ChIA–PET; YY1-HiChIP; and eQTL SNPs	106 ENCODE, 73 Roadmap, 600 FANTOM, 23 GRO-Seq
JEME (Cao et al. 2017)	Two sets: a) Roadmap ChromHMM predictions pooled across cell-types, b) FANTOM 5 enhancers	Roadmap (H3K4me1, H3K27ac, DHS, H3K27me3) and FANTOM CAGE	Multiple: LASSO and elastic net together with random forest	ChIA-PET, Hi-C and expression quantitative trait locus (eQTL)	127 Roadmap + 808 FANTOM5
GeneHancer (Fishilevich et al. 2017)	Enhancers reported in: a) Ensembl (based in DHSs and H3K27ac from Roadmap and ENCODE), b) FANTOM5, and c) VISTA db	eQTLs, eRNA co-expression, TF co-expression, CHi-C and gene target distance (nearest neighbor) combined into S_{GE} score	Score-based calculation of S_{GE}	Mendelian regulatory mutations in Genomiser, *in vivo* validated heart enhancers, literature sampling, VISTA	NA
EnhancerAtlas (Gao et al. 2016)	Enhancer defined based on 8 tracks: DHS, eRNA, P300 and POL2 binding sites, nucleosome-depletion, H3K4me1 and H3K27ac, TFBS and CHIA-PET signals	H3K4me1, H3K27ac, H3K4me3, RNA-Seq, gene-target distance	IM-PET	VISTA database (Visel et al., 2007) but only to validate enhancers	105

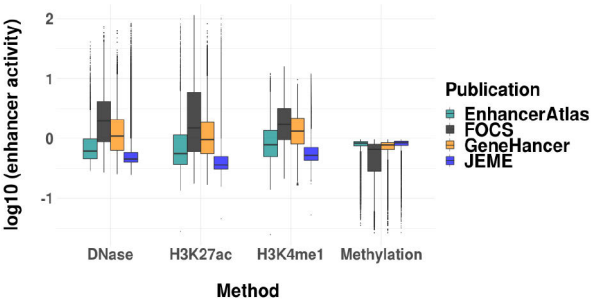
3.3.2. Enhancer definitions from different sets of EGAs vary tremendously in their properties

After we downloaded and processed enhancer-gene associations reported in JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS (Hait et al. 2018) and extracted their enhancers, we spotted the large differences in the number of reported enhancers: 1,008,243 EnhancerAtlas enhancers (out of 2 534 123 originally reported in the publication), 189,301 GeneHancer enhancers (284,834 reported), and 291,992 JEME enhancers (we pooled FANTOM5 CAGE and Roadmap enhancer regions) and 103,979 "pooled" FOCS enhancers was identified (**Supplementary Table 1.**). We additionally identified four groups of FOCS enhancers consisting of 65,423 FANTOM5, 255,925 GRO-Seq, 408,802 ENCODE and 470,546 Roadmap enhancers. However, since we additionally processed original enhancer definitions (to simplify comparisons between enhancer definitions; Methods) we consequently reduced the number of original enhancers reported in each publication; for example, 1M EnhancerAtlas enhancers was reduced to less than 200,000 enhancers. Except in their number, reported enhancers differed in the percentage of the genome they cover and their length (**Figure 3.1. A-B**).

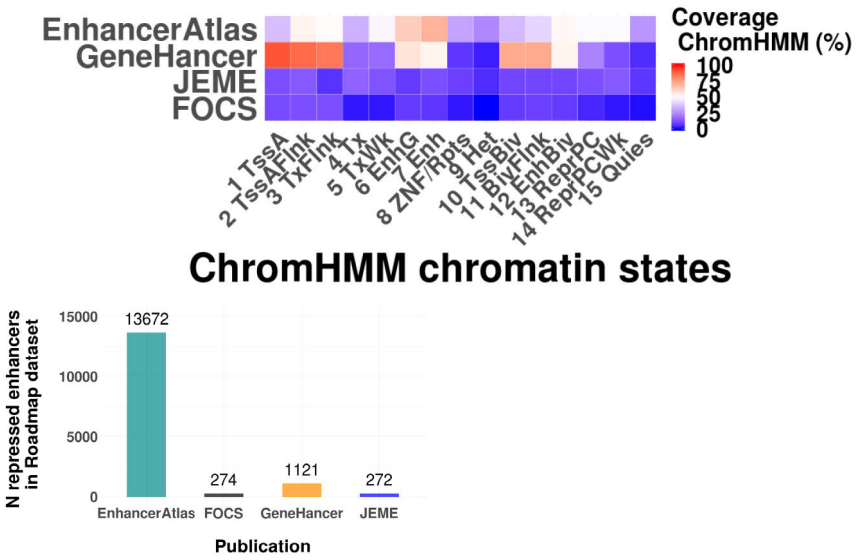

EnhancerAtlas
GeneHancer
JEME
FOCS
% of genome coverage
N of enhancers
Density
log10(enhancer length)


% annotated enhancers
introns
intergenic
exons
promoters
Genetic context
Publication
EnhancerAtlas
FOCS
GeneHancer
JEME


% of enhancer bp
Categories
Enhancer coverage category
Jaccard index (%)
EnhancerAtlas
FOCS
GeneHancer
JEME


log10 (enhancer activity)
DNase
H3K27ac
H3K4me1
Methylation
Method
Publication
EnhancerAtlas
FOCS
GeneHancer
JEME

D. **E.** **F.**



G. H.

Figure 3.1. General characteristics of enhancers reported in four publications: JEME (Cao et al., 2017), GeneHancer (Fishilevich et al., 2017), EnhancerAtlas (Gao et al., 2016), and FOCS (Hait et al., 2018). A. Percentage of the genome covered by each enhancer definition in respect with the number of enhancer regions (all overlapping enhancers were reduced to one enhancer region). The size of each point corresponds to the number of enhancers reported in the corresponding publication. B. Distribution of enhancer sizes (plotted on the log10 scale). C. Percentage of the per-base overlap between promoter, intron, exon, and intergenic regions and published enhancer regions. D. Percentage of the genome covered by enhancers (30% of the genome is covered by at least one enhancer definition); category 1 spans a number of bp that are covered by only one enhancer definition (publication-specific enhancers), whereas category 4 spans genome locations that are covered by all four enhancer definitions from different publications. E. Heatmap of calculated Jaccard index; a measure of similarity between enhancer definitions calculated as the intersection over the union. F. Distributions of the averaged enhancer activity quantified separately for each enhancer definition across all four chromatin marks (DNase, H3K27ac, H3K4me1 and DNA methylation). In short, within each cell type, the mean value of enhancer activity across all enhancers on the chromosome 1 was calculated, and then averaged across cell types. G. Averaged percentage of ChromHMM chromatin states covered by different enhancer definitions. Across each cell type, a per-base enhancer overlap was calculated for each individual chromatin state; and then it was averaged across cell types; and divided by the total length of the genome covered by an analyzed mark. H. Histogram of the number of enhancers that show full per-base overlap with regions predicted by ChromHMM as heterochromatin, repressed Polycomb, weak repressed Polycomb and quies.

Enhancer length ranged from 1 bp (GeneHancer) up to 3.2 million bp reported in EnhancerAtlas, with an average between 408.5 bp (FOCS) and 2852.4 (EnhancerAtlas; **Supplementary Table 1.**). For each publication, enhancers mostly overlap with introns and intergenic regions, and show small overlap with exons and promoters (**Figure 3.1.C**). We identified that one third of the genome is covered by at least one enhancer definition and more than two-thirds of the enhancer regions in the genome is covered by a publication specific enhancers, e.g. enhancers reported in only one publication (**Figure 3.1.D**). We calculated the Jaccard index (intersection between datasets divided by the size of the union) to measure the similarity between enhancer definitions. The highest similarity between two definitions of enhancers was observed for GeneHancer vs EnhancerAtlas enhancers (89%), whereas the lowest similarity was identified for JEME vs FOCS enhancers (17%, **Figure 3.1.E**).

For each published enhancer definition, we examined differences in the level of quantified enhancer activities. We show that enhancer activity varies across chromatin marks, cell types and enhancer definitions. Nevertheless, the rank of publication of averaged enhancer activity was preserved across chromatin marks (**Figure 3.1.F**): JEME enhancers show the lowest averaged enhancer activity signal, whereas FOCS enhancers show the highest averaged signal for three chromatin marks (H3K4me1, H3K27ac, DHS) and the lowest for DNAme.

We calculated per-base overlap between enhancers and ChromHMM-predicted chromatin states for 127 Roadmap epigenome (15-state ChromHMM model, **Figure 3.1.G**) to test whether ChromHMM predicted enhancer-like chromatin states (6_EnhG, 7_Enh, 12_EnhBiv) have a higher percentage of enhancer coverage than other states. In short, ChromHMM software learns and characterizes chromatin states by integrating multiple chromatin datasets and the resulting model can then be used to systematically annotate a genome in one or more cell types (Ernst and Kellis, 2012). Only for EnhancerAtlas enhancers we observed that the overlap between enhancer-like ChromHMM states and enhancer definitions was higher than for other chromatin states (63% overlap of 6_EnhG and 70% overlap of 7_Enh with EnhancerAtlas enhancers). However, all other ChromHMM chromatin states showed as well a high percentage of coverage by EnhancerAtlas enhancers; for example, 49% of predicted Polycomb repressed regions - 13_ReprPC and 14_ReprPCWK - were covered by EnhancerAtlas enhancers. In the case of GeneHancer, 91% of the actively transcribed regions predicted by ChromHMM (1_TssA) were covered by enhancer regions, whereas 57% and 53% of the ChromHmm predicted enhancer regions (6_EnhG and 7_Enh) were covered by the same definition of enhancers. Although all enhancer definitions should represent active enhancers in the genome, we tested whether some publication-defined enhancer regions show full per-base overlap with ChromHMM-predicted heterochromatin, repressed Polycomb, weak repressed Polycomb and quies regions. We identified 272, 274, 1,121, and 13,672 such enhancers for JEME, FOCS, GeneHancer and EnhancerAtlas.

3.3.3. Comparing enhancer-gene associations (EGAs) across computational datasets

We additionally compared enhancer-gene associations across different methods. We identified 117 355, 506 471, 929 682, and 2 327 946 enhancer - gene associations in FOCS, GeneHancer, JEME and EnhancerAtlas, respectively (**Figure 3.2.A**, **Supplementary Table 1.**).

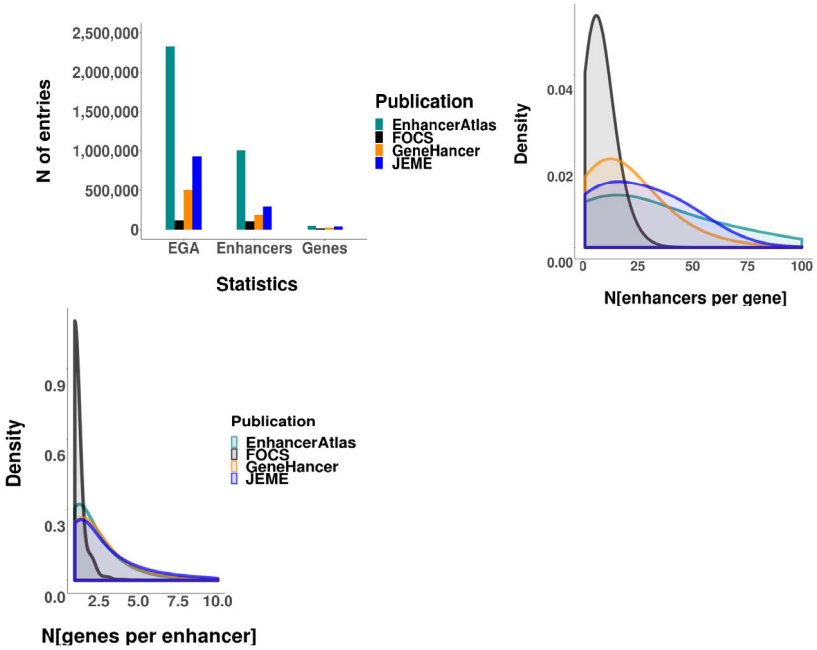


Figure 3.2. General characteristics of enhancer-gene associations (EGA) reported in four publications: JEME (Cao et al., 2017), GeneHancer (Fishilevich et al., 2017), EnhancerAtlas (Gao et al., 2016), and FOCS (Hait et al., 2018). A. Number of enhancer-gene associations, enhancer and genes reported in each study. B. Per publication distribution of the number of genes associated with each enhancer region. C. Per publication distribution of the number of enhancers reported for each gene.

FOCS had, on average, less associated genes per enhancer and enhancers per gene (median ranged between 6 for FOCS up to 24 for EnhancerAtlas (**Figure 3.2.B)**. A maximum of 608 enhancers for one gene was reported in the EnhancerAtlas (**Supplementary Table 1.**). One enhancer region was associated with a maximum of 52 genes in GeneHancer and EnhancerAtlas reported EGAs.

To investigate the publication-specificity of reported enhancer-gene associations, we overlapped enhancer - gene associations reported in different publications as described in Methods. The highest number of overlaps between two sets of EGAs was identified for GeneHancer-FOCS EGAs; 59% of GeneHancer EGAs was confirmed by FOCS EGAs (59%, **Figure 3.3.**), whereas 8% of FOCS EGAs is confirmed by GeneHancer EGAs. In addition, only 3% of FOCS EGAs were confirmed by JEME and 7% by EnhancerAtlas.

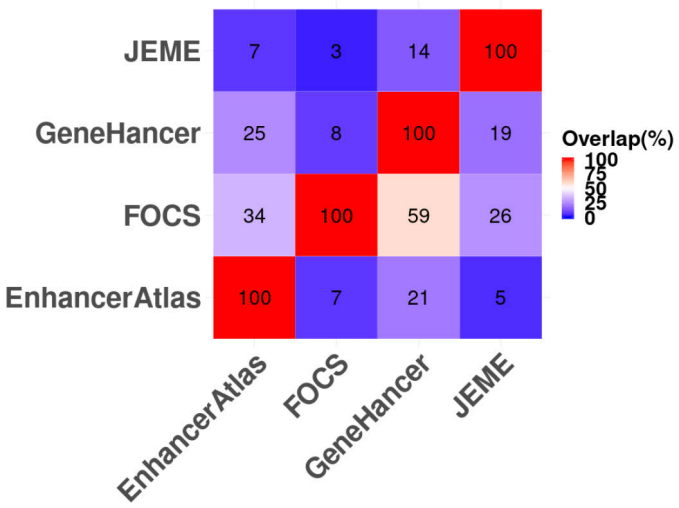


Figure 3.3. The heatmap of the percentage of overlaps between sets of enhancer-gene associations (EGAs). Percentage was calculated for each set individually, and it corresponds to the number of identified overlaps divided by the total number of EGAs in analyzed publication. Columns correspond to the queried datasets: for example, 59% of GeneHancer EGAs is "confirmed" by FOCS EGAs, whereas 8% of FOCS EGAs is confirmed by GeneHancer EGAs.

3.4. Discussion

We identified multiple computational methods that study enhancer-gene associations (He et al. 2014; Corradin et al. 2014; Roy et al. 2015; Gao et al. 2016; Whalen et al. 2016; Cao et al. 2017; Fishilevich et al. 2017; O'Connor et al. 2017; Hait et al. 2018; Li et al. 2018; Libbrecht et al. 2018; Wang et al. 2019, Singh et al. 2019). However, we thoroughly reviewed four methods: JEME - joint effect of multiple enhancers (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS - FDR-corrected OLS with Cross-validation and Shrinkage (Hait et al. 2018), because we were primarily interested in analyzing publications that had a computational component, which models were built upon a large number of cell lines and that predicted several thousands interactions. This is the first time that results and protocols of different computational methods were systematically compared and we corroborated previous observations that the initial choice of epigenomic marks used to map enhancers has a large impact on the final number and characteristics of defined enhancers (Zentner and Scacheri 2012), and consequently, on the number and properties of defined EGAs (Hariprakash and Ferrari 2019).

In terms of enhancers, publications differed more than a ten-fold in their number: from hundred thousand enhancers reported in FOCS to up to one million unique enhancers in EnhancerAtlas 2.0. Importantly, some publications (JEME and EnhancerAtlas) predicted enhancers and EGAs individually, for each analyzed cell type/tissue, whereas others (GeneHancer and FOCS) reported no cell-type specific predictions, and thus, to enable comparison of enhancer regions, we pooled enhancers predicted in JEME and EnhancerAtlas across all cell types and processed them additionally. This caused the reduction in the total number of enhancers: 1M EnhancerAtlas enhancers was reduced to less than 200,000 enhancers. Nevertheless, the difference in the number of enhancers reported across methods remained.

A discrepancy in the number of predicted enhancers in the human genome has been previously reported in the literature: an initial estimation of 1 million enhancer regions (Heintzman et al. 2009, ENCODE Project Consortium 2012), was exceeded when by Thurman et al. (2012) predicted 2.9 predicted million DHSs many of which are likely enhancer regions, or downsampled to 43,011 enhancer candidates based on FANTOM5 CAGE datasets (Andersson et al. 2014). However, even today, we still do not know the true number of enhancers and we lack a genome-wide exhaustive reference list of all non-coding regions that can act as enhancers (Hariprakash and Ferrari 2019). Especially, since none of the currently discovered chromatin marks was found to be perfectly

predictive of an enhancer activity (Shlyueva et al. 2014) and the number of newly discovered marks associated with an enhancer activity is still growing (Stricker et al. 2017; Mathelier et al. 2015). This is further complicated by their dynamic and cell type-specific nature of enhancers (Joshi 2014, Chen et al. 2018) and the fact that enhancer relative location, with respect to its target genes, can be greatly variable (Lettice et al. 2003).

Thus, different publications used different protocols to define enhancers. In general, FOCS and JEME enhancer regions were defined as an aggregate of enhancer definitions across different data types and/or sources, whereas GeneHancer and EnhancerAtlas represent a weighted consensus of different enhancer definitions. Specifically, FOCS used peaks from various GRO-Seq, FANTOM5 CAGE, Roadmap DHSs, ENCODE DHSs tracks to define enhancers, whereas EnhancerAtlas integrated eight different data types such as information about P300 and POL2 binding sites, CHIA-PET signals, levels of histone modifications, etc. On the other hand, Fishilevich et al. simply pooled previously reported enhancers from the four data sources: ENCODE (ENCODE Project Consortium 2012), the Ensembl regulatory build (Cunningham et al. 2015), FANTOM5 (Andersson et al. 2014) and the VISTA Enhancer Browser (Visel et al. 2007) to define GeneHancer enhancers. In JEME, Cao et al. (2017) used ChromHMM-based 15-state (Ernst and Kellis 2012) systematic annotations of 127 Roadmap cell types and pooled and processed the identified regions to define a final set of enhancers.

Upon identifying differences in their number, we set off to investigate other properties of defined enhancers. We showed that reported enhancers differed in the percentage of the genome they cover, their length, and coverage of other functional elements in the genome such as promoters, intronic and exonic regions. For example, JEME enhancers had lowest overlap with promoter regions, whereas FOCS had the highest overlap with introns. Although regulatory elements may have both enhancer and promoter functions (Andersson et al. 2015; Andersson and Sandelin 2020), the percentage of promotes that have a strong enhancer activity in the genome was shown to not be large (3% out of 20,719 tested gene promoters in human K562 cells had strong enhancer activity *in vitro,* Dao et al. 2017), and we expected to observe that the highest number of enhancer regions overlap the non-coding introns of genes or intergenic regions. Especially since intronic enhancers were commonly found to be engaged in the long-range gene interactions with distant genes (Pomerantz et al. 2009; Harismendy et al. 2011; Maurano et al. 2012; Smemo et al. 2014). Indeed, this was the case for all methods.

On the other hand, sets of enhancers did not necessarily show the highest overlap with ChromHMM predicted enhancer-like chromatin states (6_EnhG, 7_Enh, 12_EnhBiv, Ernst and Kellis 2012) across 127 Roadmap cell types and tissues (Roadmap Epigenomics Consortium et al. 2015). In short, ChromHMM software was used to learn chromatin states by integrating multiple chromatin datasets and systematically annotate 127 Roadmap genomes into 15 chromatin states such as: enhancers, active transcribed regions, Polycomb repressed regions, etc., and we expected to observe that computational predicted enhancers show the highest coverage with three ChromHMM-predicted enhancer-like chromatin states. However, we observed this hold true only for EnhancerAtlas enhancers, which however, showed high overlap with other, non-enhancer ChromHMM-predicted chromatin states. Unfortunately, this is likely a consequence of the fact that EnhancerAtlas reported enhancer regions that are, on average, longer than enhancer regions defined in other publications and not mutually exclusive. Strikingly, some of defined enhancers are even located in "inactive" chromatin states (as predicted by ChromHMM as heterochromatin, repressed Polycomb, weak repressed Polycomb and quies regions), across all 127 Roadmap cell types (Roadmap Epigenomics Consortium et al. 2015). Strikingly, because we would assume that such regions should not be predicted to be enhancer regions, especially since most of the methods used Roadmap datasets as an input for their algorithms.

Quantified enhancer activity varied across enhancer definition as well; this further indicated that enhancers cover distinct regions in the genome. To corroborate this observation, we performed an additional analysis that demonstrated that two-thirds of the enhancer regions in the genome were covered by a publication specific enhancers, e.g. enhancers reported in only one publication, whereas, the two methods that, on average, reported the largest enhancers (GeneHancer and EnhancerAtlas) exhibited the biggest overlap between enhancer definitions. It is important to emphasize that one third of the genome was actually reported to be an enhancer by at least one of the analyzed methods, as compared to the previous report by (Li et al. 2018) that labelled 6.8% of the genome as enhancers and 0.6% as promoters in one or more of six well-characterized cells.

Since we were further interested in pursuing computational modelling of *enhancer activity ~ gene expression*, this analysis posed a grim question, which, if any, of enhancer definitions should we use in our future research or functional analyses of risk associated SNPs. This will be answered in the future chapters of this thesis.

We showed that the diversity of enhancer definitions had a large influence on the final predictions of enhancer-gene associations (EGAs). By identifying a large discrepancy in the number and properties of identified sets of EGAs, we corroborated the previous observations by Hariprakash and Ferrari 2019 that the first practical problem of computational methods that study *cis* regulatory interactions is actually an identification of enhancer regions themselves. However, to be able to compare results coming from different methods, we had to perform an extensive data integration. As compared to predictions of enhancer regions, the complexity of EGA predictions is even more puzzling; it is complicated by the fact that enhancers tend to "skip" over a proximal to gene to regulate a more distant one (Merli et al. 1996), regulate more than one gene (Simonet et al. 1991; Schwartz and Olson 1999; Bender et al. 2001), and activation of a specific gene may be activated in multiple cell types by distinct enhancers (Burch 2005; Abbasi et al. 2007; Landry et al. 2009; Visel et al. 2009). Likewise, since the selection of computational algorithms and data sources to assess E-G associations differed across methods, which should, as well, have a great impact on the final enhancer-gene predictions. Especially since each method (either experimental or computational) suffers from specific or general technical and biological limitations of HTS technologies and can additionally hinder the creation of an exhaustive reference list of enhancers.

3.5. Conclusions

In general, enhancer-gene associations coming from different methods are hardly directly comparable. Due to the differences in the way multiple parameters and information are used to define enhancers and promoters itself, algorithmic details, but also, how true positive ETG pairs are defined, final sets of assessed enhancer-gene associations differ. With this in mind, we set off to develop and benchmark a computational method that maps *cis*-regulatory interactions in the genome, but at the same time integrates results of the previous methods.

4

Results II: reg2gene - a novel computational method to associate enhancers and genes

Preface

In this section, I set off to develop a novel computational method that associates regulatory regions with genes they regulate in a genome-wide manner - the reg2gene algorithm. Method was designed together with Dr. Altuna Akalin and Dr. Vedran Franke.

Abstract

We develop a novel computational method that associates regulatory regions with genes they regulate in a genome-wide manner - the reg2gene algorithm.

reg2gene was built upon extensive data modeling and integration, and as such, it consists of three main steps: 1) data quantification, 2) data modelling and significance assessment, and 3) data integration. reg2gene relies on the largest collection of epigenomics or transcriptomic data in humans at time - the Roadmap datasets and its five epigenomics subsets: H3K4me1, H3K27ac, DNAme, DHS and RNA-Seq; and implements three correlation-based methods: Pearson, Spearman and distance correlation and executes elastic net regression and random forest. Models were integrated by the majority voting approach and further improved by an ensemble voting with enhancer-gene associations (EGAs) reported in: EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), and FOCS (Hait et al. 2018). As a result, we report two sets of enhancer-gene associations: a flexible set of ~230K EGAs reported in at least 2 publications) and a stringent set of ~60 EGAs reported in three or more publications. Predicted sets of EGAs generally colocalized within the same topologically associating domain.

reg2gene functions were gathered in the reg2gene R package. This package can enable solving of similar biological problems such as modelling across cell types, tissues or time course series.

4.1. Introduction

Most regulation of gene expression is believed to occur at the level of transcription initiation regulated by *cis*-regulatory sequences in the genome, such as enhancers and promoters, that recruit a distinct set of *trans* factors (Levine and Tjian, 2003). Although certain aspects of gene regulation are understood, the precise methodology for its research is missing and the field lacks an approach that is systematic, integrative and accessible for discovering and documenting *cis*-regulatory relationships across the genome.

Different experimental and computational techniques have been in use to study enhancer-mediated gene expression regulation in a genome-wide manner and they can be broadly categorized into four categories: computational modelling of *gene expression~enhancer activity* (Hariprakash and Ferrari, 2019), eQTL studies (GTEx Consortium et al., 2017), HiC technologies (Lieberman-Aiden et al., 2009), and direct functional confirmation of enhancer activity by reporter assays or cellular screens (Arnold et al., 2013; Kheradpour et al., 2013; Kvon, 2015; Fulco et al., 2019; Gasperini et al., 2019).

Since we previously showed that individual sets of predictions differ tremendously, especially in the location, number and properties of defined enhancer regions and enhancer-gene associations, we set off to develop a novel computational method that associates regulatory regions with their gene targets in a genome-wide manner, but that, at the same time, overcomes shortcomings of the previous computational methods.

4.2. Methods

4.2.1. The intuition behind the voting procedure

Majority voting is a simple binary decision rule that opts for an alternative that has a majority, that is, more than half the votes. By creating an ensemble classifier one can combine the classification rules of multiple classifiers and potentially narrow down the hypothesis space (**Figure 4.1.**). Such an ensemble, created by averaging or majority voting often appears to be more accurate than individual predictions (Yang et al., 2010).

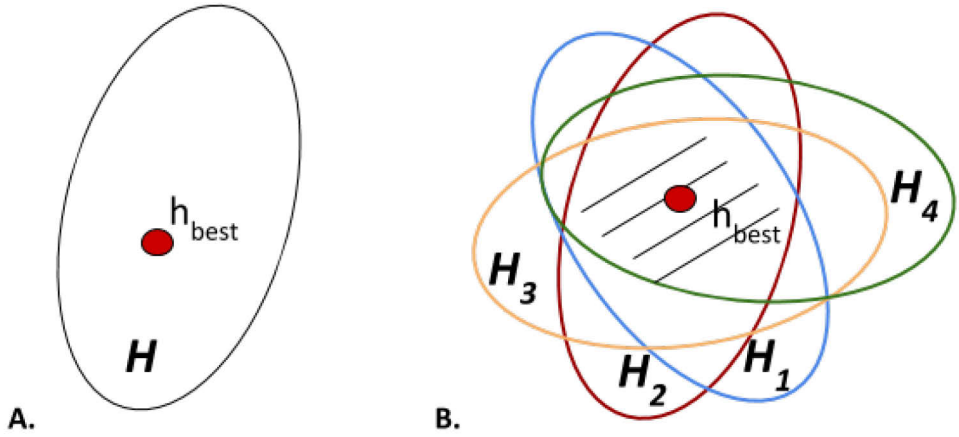


Figure 4.1. A schematic illustration of partitioning and narrowing a hypothesis space by using an ensemble classifier. (A) Hypothesis space of a single classifier, (B) Hypothesis space of an ensemble classifier.

Nonetheless, in order to improve predictions, the base classifiers need to be accurate (better than chance) and diverse from each other (Tsymbal et al., 2005). This need for diversity originates from the assumption that if a classifier makes a misclassification there should be another classifier that complements the first classifier by correctly classifying the misclassified sample. Ideally, each classifier makes incorrect classification independently.

4.2.2. An overview of used datasets

I used several publicly available datasets and original publications to define locations of enhancers and genes, predict enhancer-gene associations, and improve and benchmark predictions of EGAs (**Table 4.1.**). To achieve that I performed data integration on various levels by combining information from:

1) epigenomes profiles (histone modification patterns, DNA accessibility, and DNA methylation from the Roadmap dataset; (Bernstein et al., 2010; Roadmap Epigenomics Consortium et al., 2015),

2) computational algorithms that model gene expression ~ enhancer activity,

3) computational methods developed to predict enhancer-gene associations: JEME, EnhancerAtlas, GeneHancers, FOCS (Cao et al., 2017; Fishilevich et al., 2017; Gao et al., 2016; Hait et al., 2018).

Since I already reviewed enhancer-gene associations reported in JEME, EnhancerAtlas, GeneHancers, FOCS, here, I will only review datasets that I used to define enhancers and genes and model their activity. Nonetheless,those datasets are thoroughly described in Chapter 2 - Methods.

Table 4.1. Characteristics and statistics behind used datasets in this chapter

Data source	Technology/Data type	Processing algorithm	Used as/for:	N of EGAs/interactions
NIH Roadmap	ChIP-Seq for H3K27ac, H3K4me1, BS Seq for DNA methylation and DHS Seq	NA	Prediction of enhancer-gene associations	NA
ChromHMM processed NIH Roadmap	H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3	multivariate Hidden Markov Model	Defining enhancer regions	NA

GENCODE (Harrow et al., 2012)	RNA-Seq, RT-PCR, RACE	multiple	Define gene regions	NA

4.3. Results

4.3.1. The reg2gene algorithm

reg2gene consists of the three main steps that are executed in the following order:

1) data quantification,
2) data modelling and assessment, and
3) filtering and voting (**Figure 4.2.**).

Step 1. Data quantification:

Quantifying enhancer activity with regActivity function from the reg2gene R package

For each enhancer region, reg2gene individually quantifies its activity using .bigwig files as an input. For example, bigwig files for ChIP-Seq, DNase-Seq or WGBS signals. It calculates the mean coverage value of the input signal (for example H3K4me1 ChIP-Seq signal) over each enhancer location and quantile normalizes quantified signals. This procedure is implemented as a regActivity() function and for example, we ran it four times to quantify enhancer activity based on four chromatin marks.

Quantifying gene expression with bwToGeneExp function from the reg2gene R package

For each gene, gene expression levels are quantified based on the signal from the genome-wide RNA-Seq coverage tracks. First, for each exon region, the mean coverage value of the input signal is calculated and quantile normalized as the log(gene expression scores + 1). Then, a sum of the mean exon expressions (mean exon expression multiplied by exon length) is divided by a full gene length (a sum of the exon lengths).

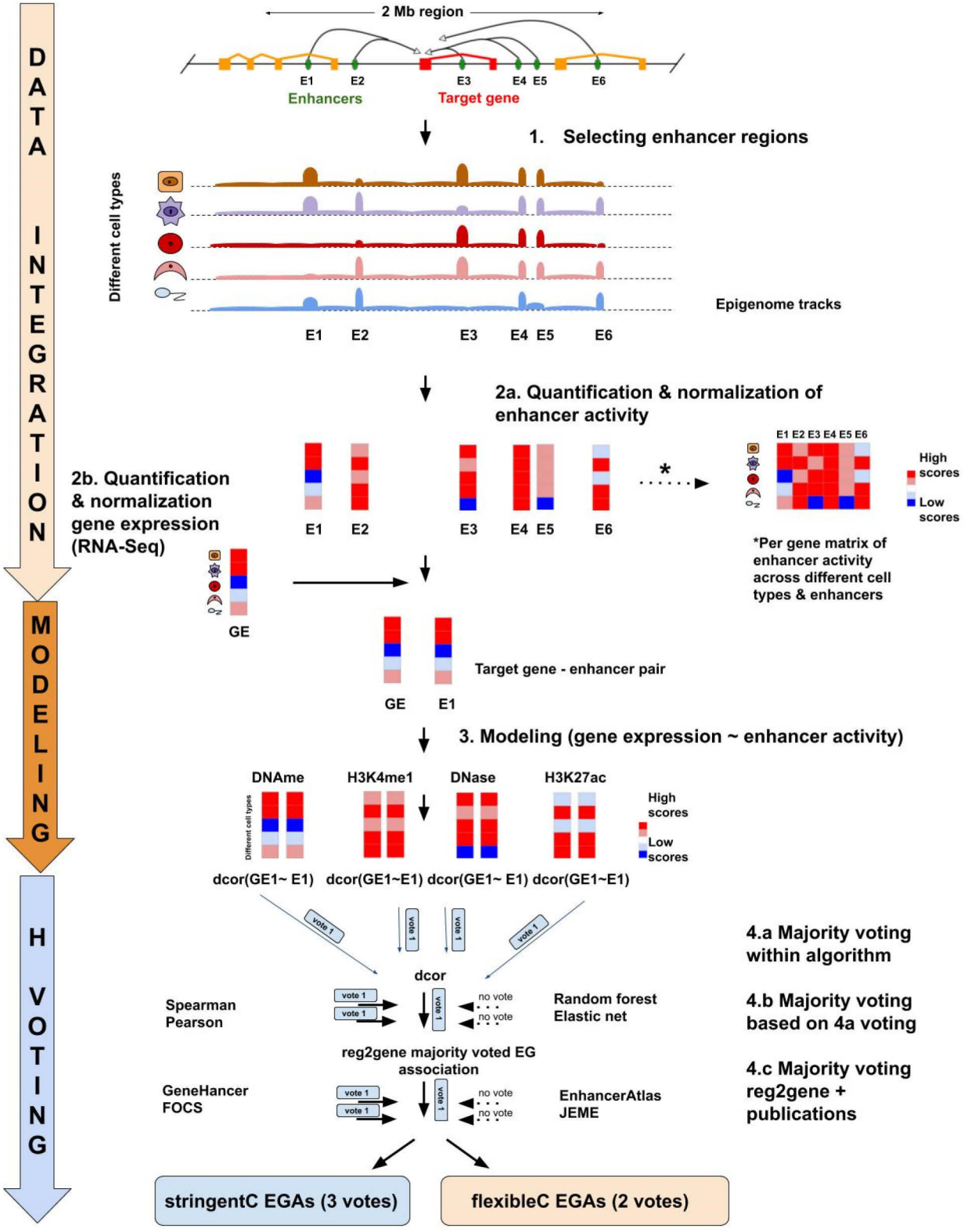


Figure 4.2. A workflow of reg2gene modelling procedure. A three-step design is described in detail in this Chapter.

Step 2. Modelling and assessment

Modelling

For each gene, enhancers that are located within +/-1Mb from the transcription start site (TSS) are identified (using the regActivityAroundTSS function) and gene expression~enhancer activity models are built using five algorithms: Pearson and Spearman correlation coefficient, distance correlation, elastic net and random forests. Before modelling, enhancer activity and gene expression scores were scaled and centralized. Elastic net modelling is used together with five-fold cross-validation (cv.glmnet() and lambda that gives minimum mean cross-validated error (lambda.min) was selected. Random forest was run with 500 trees with the Gini index as a measure of variable importance. For each model, associateReg2Gene function reports association statistics - it estimates p-values for reported coefficients by randomly resampling response variables (gene expression) 1000 times, and from the resampling statistics it assesses P-value based on the Gamma distribution. In addition, the algorithm implements data scaling and centalizes data.

Assessment of statistical significance

For each combination of algorithms and chromatin marks, reg2gene individually performs an assessment of the model significance across the full set of modelled enhancer-gene pairs: it simultaneously threshold models by its q-value and modelling statistics using predefined threshold level.

The rationale behind calculating the model-specific threshold (**Figure 4.3.**) is that, contrary to the q-value which can be uniformly used as a threshold across all combinations of algorithms and chromatin marks, such a single value threshold could not be applied for modelling statistics which was individually calculated by five different algorithms. For example, Pearson and Spearman correlation coefficient has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. Distance correlation coefficient is 0 if and only if the random vectors are independent, whereas +1 describes perfectly dependent variables.

On the other hand, elastic net and random forest work on the rationale that multiple enhancers can regulate a single gene, and thus, they identify significant relationships between multiple enhancers and target genes simultaneously as they assess the strength of impact of those enhancers on their target. Random forest variable importance is measured by the variance of responses and is only limited by the minimum lower value of zero. For example, the maximum value across all reg2gene random forest models for the variable importance was 74. Elastic net coefficients calculated for each enhancer-gene pair have no predefined upper and lower limit since they were calculated using five-fold cross-validation procedure and assessed for lambda that minimizes mean cross-validated error. Thus, since each of the five algorithms calculated enhancer-gene association differently, we decided to establish a procedure that will uniformly identify model-specific thresholds.

Statistical significance assessment protocol

Due to its limited upper and lower value, Pearson and Spearman correlation coefficient is the most intuitive measure of statistics: +/-0.3 is usually considered to imply a weak positive/negative correlation; +/-0.5 moderate positive/negative correlation; whereas +/-0.7 implies strong positive/negative correlation. Thus, we decided to develop a uniform thresholding procedure based on results of Pearson and Spearman correlation models.

First, all models that report Pearson and Spearman correlation coefficient are analyzed. We sorted model statistics within each model type and identified a value that corresponds to the third quartile of model statistics (**Figure 4.2.**). We assess this split based on the observation that roughly one-quarter three-quarter split in each of analyzed Pearson or Spearman-based model series had an absolute correlation coefficient above 0.3.

The inferred value of the third quartile was used to assess model significance - statistics above the assessed value corresponds to the significant results, whereas all values below the threshold were considered not to be significant.

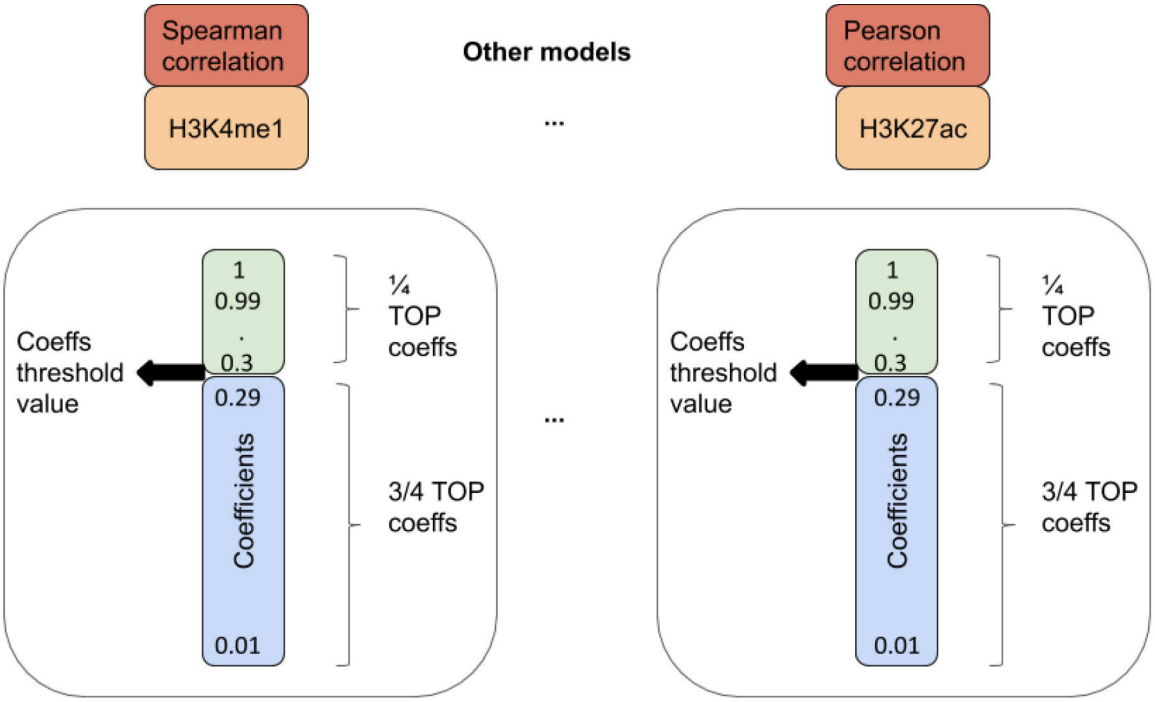


Figure 4.3. Schematic representation of the model significance assessment. For each out of twenty models (for example, one model corresponds to Spearman correlation coefficients for 8.9M enhancer-gene pairs that were assessed based on the enhancer activity quantified using H3K4me1 signals) assessment was performed individually as described in text. In short, model statistics that were found to be above the certain value - the measure of the third quartile - was considered significant, whereas all values below that threshold were considered not to be significant.

Step 3. Filtering and hierarchical majority voting

pre-voting filtering

Since reg2gene modelling procedure will inevitably fail to model certain associations, simply because it relies on a scarce dataset(s), we developed a filtering protocol that filters out such genes and enhancer regions. Thus models with:

a) zero standard deviation of gene expression or enhancer activity across different cell types,

b) more than 90% of the cell types with the expression level or enhancer activity equal to zero;

c) at least one across-cell-type overrepresented value (maximum of 10 cell types can share the same value);

d) characterized by a small number of gene expression or enhancer activity clusters which is an indicator of a bad imputation protocol (a minimum 30 unique values, across different cell types, was requested to be quantified for each gene).

In addition, a stepwise procedure was implemented to test and remove missing data:

1) cell types that had missing information for more than 75% of tested enhancers;
2) enhancers/genes with missing activity level in more than 75% of cell types,
3) the remaining missing values.

Hierarchical majority voting

reg2gene performs hierarchical majority voting to integrate information from different algorithms and epigenomic marks and assess the final modelling success. As mentioned in the introduction, majority voting is a simple binary decision rule that opts for an alternative that has a majority, or in other words, more than half the votes. Thus, for each analyzed enhancer-gene link reg2gene integrates twenty binary scores obtained by modelling and assessing the statistical significance via 2-step voting procedure (**Figure 4.4.**).

Step 1: For each enhancer-gene pair, we counted the number of significant associations assessed **within each algorithm or across chromatin marks**. In our case, there were a total of four models (corresponding to four epigenomic marks used to quantify enhancer activity) ran using one computational algorithm, and thus by statistical significance assessment each EG pair could achieve a score from zero to four for the voted algorithms. For example, voting was performed based on results of modelling using distance correlation models for H3K4me1, H3K27ac, DNA methylation and DHS datasets. If more than two votes were identified (e.g. this EG pair was assessed to be significant by more than one chromatin mark) such enhance-gene pair was considered to be voted (by an algorithm) using the majority voting approach. We repeated this voting analysis individually for each of the five algorithms.

Step 2: The majority voting approach is repeated, but across step 1 voting results (**voting across algorithms**). In this case, each enhancer-gene pair could achieve five votes for five algorithms. Enhancer-gene pairs that achieved more than three votes in the second step of the voting procedure were considered to be hierarchically majority voted.

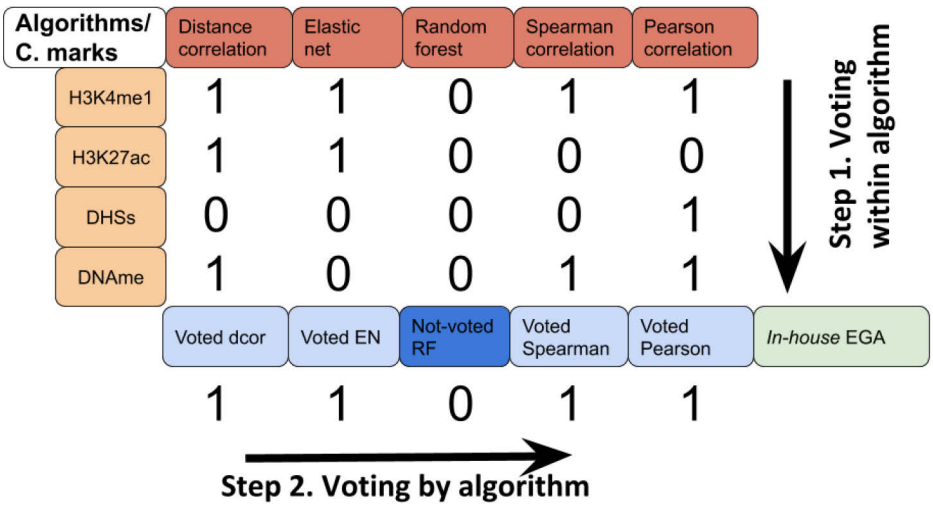


Figure 4.4. Schematic representation of the hierarchical voting procedure, We first voted models assessed with one algorithm and then voted them across assessed algorithmic votes. There were a total of four models (H3K4me1, H3K27ac, DNA methylation and DHS epigenomic datasets used to quantify enhancer activity) for each computational algorithm, and thus, by statistical significance assessment each EG pair could achieve a score from zero to four when vote within algorithms, and 0-5 when voted across step 2 voting results.

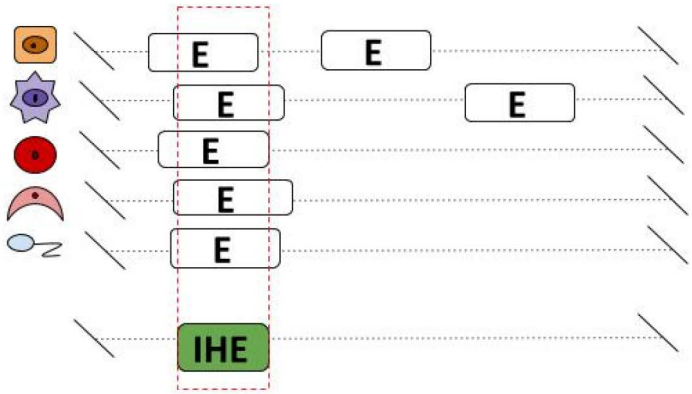
Step 3 (optional): voting with previous methods by defining a union of all associations across EGA methods.

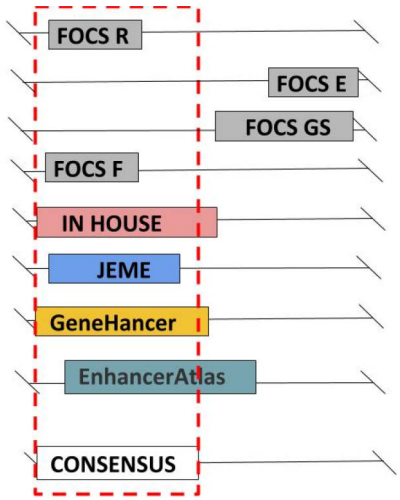
4.3.2. reg2gene R package

reg2gene is available as an R package at the Github: https://github.com/BIMSBbioinfo/reg2gene

4.3.3. More than 280 thousands enhancer regions are predicted by three or more enhancer definitions

Since we showed that the majority of previously reported enhancers are publication-specific, we decided to define "consensus" enhancer regions and use them for computational modelling. We "recycle" the idea that the majority voting approach can improve the accuracy of computational models (Yang et al., 2010) and treated sets of enhancers (H1 - EnhancerAtlas, H2 - GeneHancer, H3 - JEME , H4 - FOCS, H5 - in-house enhancers) as classifiers that define individual hypothesis spaces and which integration narrows down the hypothesis space and return more accurate predictions. We first defined in-house enhancers as explained in Methods (**Figure 4.5.A**), and then, we voted them together with other enhancer definitions (**Figure 4.5.B**). We identified a total of 184,005 "in-house" enhancer regions and 286,723 "consensus" enhancers (*consensusE*).





A. **B.**

Figure 4.5. Schematic representation of defining in-house (IHEs) and "consensus" enhancers - *consensusE*. A. We defined IHE as follows: we screened 127 Roadmap cell-types for ChromHMM-predicted chromatin states and identified regions in the genome that were predicted as EnhG_6 and Enh_7 in more than 25 cell types. We processed

identified enhancer regions as described in Methods. B. We screened across 8 enhancer definitions and identified regions that were reported as enhancers in at least three datasets. Enhancer overlap was identified on the level of a base pair.

We showed that *consensusE* have constrained sizes, less likely to overlap gene promoters and more likely reside in introns and intergenic regions as compared to other publication-reported enhancers (**Figure 4.6.**). For example, less than 5% of in-house enhancers overlap promoters, as compared to 19.7% GeneHancer enhancers. Across all methods, the highest overlap of enhancers is identified for intron regions.

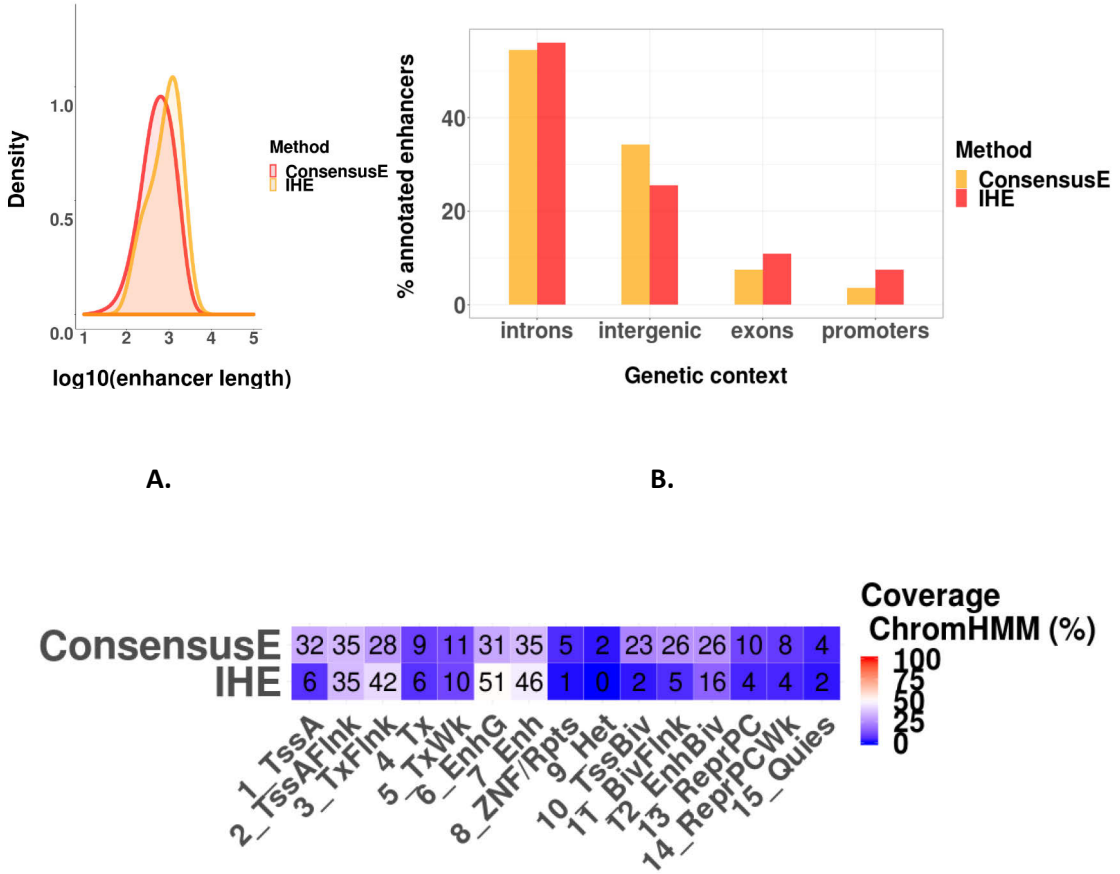


Figure 4.6. General characteristics of the two sets of defined enhancers: in-house enhancers (IHE) and *consensusE*. A. Distribution of enhancer lengths (plotted on the log10 scale). B. Percentage of the per-base overlap between promoter, intron, exon, and intergenic regions and defined enhancer regions. C. Histogram of the percentage of ChromHMM chromatin states covered by different enhancer definitions averaged across cell types. For each cell type, a per-base enhancer overlap was calculated for each individual chromatin state; averaged across cell types; and divided by the total length of the genome covered by an analyzed mark.

4.3.4. reg2gene modelling and voting identified sets of "consensus" E-G associations

After we identified "consensus" enhancer regions, we used those regions as an input for the reg2gene algorithm. First, their activity was quantified four times - based on the four imputed Roadmap datasets for different chromatin marks: H3K4me1, H3K27ac, DNA methylation and DHS. Next, gene expression was quantified for each of 34,359 GENCODE genes (Harrow et al., 2012). To test the precision of the reg2gene quantification protocol by comparing previously reported RPKM scores and reg2gene quantified values across 127 Roadmap cell types (in details described in Method section). We observed that the Pearson correlation coefficient between reported RPKMs and quantified values was consistently high across cell types (range: 0.82-0.98, median=0.95, mean=0.943).

We paired enhancers and genes that were located within +/- 1Mb window together and identified nearly 8.9 million enhancer - gene pairs (**Chapter 2**). We used those pairs and their quantified signals as an input for the modelling step of the reg2gene algorithm - we modelled *gene expression~enhancer activity* using Pearson and Spearman correlation coefficient, distance correlation, elastic net and random forest. On average, we modelled 267 enhancers per individual gene (range: 1-574, **Supplementary Table 1.**), and 25 genes per enhancer (range: 1-150).

For each enhancer-gene pair a total of 20 modelling statistics was produced (enhancer activity was quantified based on levels of four chromatin marks, and each of them was modelled using five algorithms) that were thresholded to assess the statistical significance (as explained previously). In a nutshell, each of the models (for example enhancer activity (H3K4me1)~gene expression modelled using elastic net) was thresholded by requesting: 1) q-value to be lower than 0.1 and 2) the absolute value of model statistics to be higher than model-specific threshold inferred (as explained above). The identified binary score reported whether a given enhancer-gene pair is statistically associated or not. By integrating multiple types of information across all twenty models via previously explained hierarchical majority voting procedure (in the step1 we performed voting across chromatin marks and then we voted across step1 results), for each enhancer-gene pair we identified one final binary score that reported whether or not that pair was statistically associated. We identified 1,007,448 such enhancer-gene associations and

referred to them as an "in-house" model (***inhouseM***). For this set of EGAs, a median of three genes per enhancer element was identified (mean=5.5, max=96), and 25 enhancers per gene (max=172, **Supplementary Table 1.**).

To improve the robustness of our predictions, we further voted *inhouseM* together with enhancer - gene associations predicted by JEME, GeneHancer, EnhancerAtlas and FOCS methods (Cao et al. 2017, Gao et al. 2016, Fishilevich et al. 2017, Hait et al. 2018). We identified associations which overlapped with other EGAs at least twice (***flexibleC***) or three times (***stringentC*** consensus enhancer - gene associations). Both datasets, the *stringentC* and *flexibleC*, span several thousands enhancers-gene associations: 227,271 and 61,240, respectively (**Figure 4.7.A**). *flexibleC* dataset spans 21,124 genes and 107,489 enhancers, has a mean of 1.94 genes per enhancer and a median of 9 for enhancers per gene (mean=12.4). On average, stringentC models count 14,225 genes and 44,559 enhancers (mean of genes per enhancer=1.25; median [enhancers per gene]=3, mean=4.2, **Figure 4.7.C-B, Supplementary Table 1.)**.

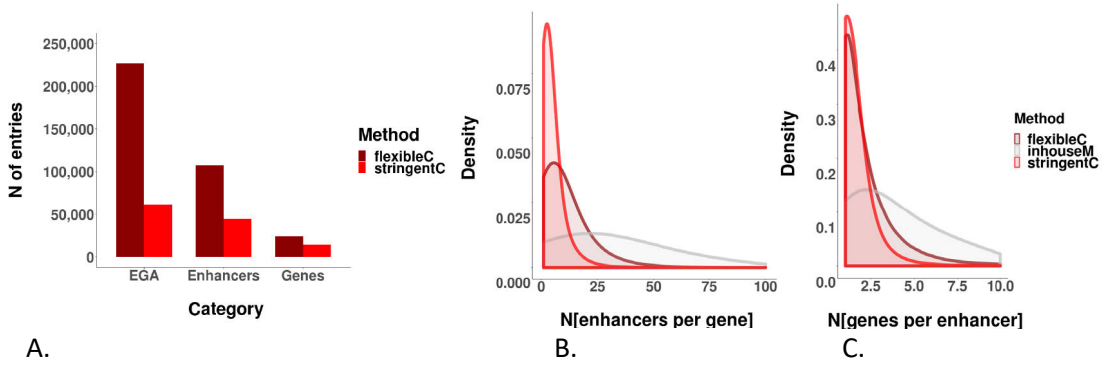


Figure 4.7. General characteristics of in-house (*in-house*) and consensus enhancer-gene associations (EGA): *flexibleC* and *stringentC*. A. Number of enhancer-gene associations, enhancers and genes reported in *flexibleC* and *stringentC* predictions. B. Per method distribution of the number of genes associated with each enhancer region. C. Per method distribution of the number of enhancers reported for each gene.

4.3.5. The majority of EGAs co-localize within the same TAD

Since enhancers and their target genes have been frequently found to co-localize in the same TAD (Symmons et al. 2014) and spatial organization of the genome represents another level of gene expression regulation (Spielmann et al. 2012; Spielmann and Mundlos 2013; Ibn-Salem et al. 2014;

Lupiáñez et al. 2015), we hypothesized that the colocalization profile of EGAs within TADs can be used to assess the performance of EGA predictions.

To test that, we calculated and averaged the percentage of enhancer-gene associations that were located within the same TAD in 37 cell types/tissues from Lieberman-Aiden et al. 2009, ENCODE Project Consortium 2012,Rao et al. 2014, Dixon et al. 2015, Leung et al. 2015, Schmitt et al. 2016. Specifically, for each cell type and set of enhancer-gene associations, we counted the number of EGAs that are located within the same TAD regions and calculated the summary statistics (**Figure 4.8.**).

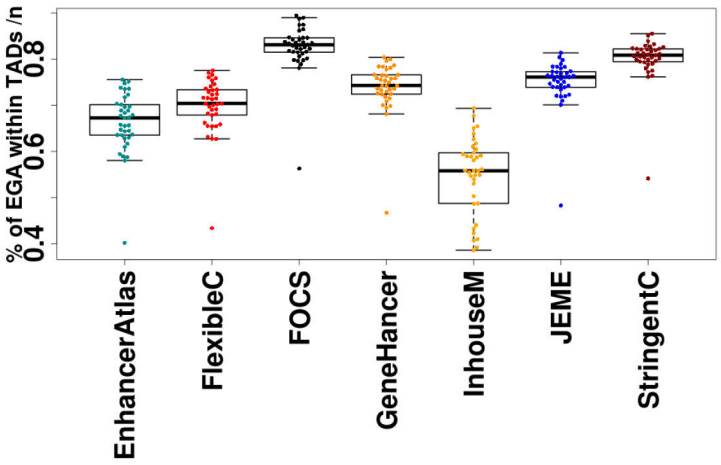


Figure 4.8. Distribution of the percentage of enhancer-gene associations (EGAs) that were found to be located within the same TAD domain tested across 37 cell types from Lieberman-Aiden et al. 2009, ENCODE Project Consortium 2012,Rao et al. 2014, Dixon et al. 2015, Leung et al. 2015, Schmitt et al. 2016.

Sets of the FOCS and *stringentC* reported EGAs showed the best summary statistics - percentage of EGAs located within the same TAD across cell types (median=83.5% [range: 58.2-90.1%] and median=81.6% [56.8-86.6%]). On the other hand, inhouseM EGAs had on average the lowest percentage of colocalization within TAD regions (median=54.1% [range: 27.9-69.3%]).

4.4. Discussion

reg2gene associated genes with their regulatory regions based on assessment of correlation between enhancer activity status and its target gene expression across multiple cell types. This approach was based on the observation that regulatory information can be particularly revealing when compared across many cells or tissues within a single genome (Ernst et al. 2011, Sheffield et al. 2013, Varley et al. 2013, Reuter et al. 2015). Previously, Ernst et al. (2011) correlated normalized signal intensities of H3K27ac, H3K4me1 and H3K4me2 with gene expression across nine cell types, whereas Sheffield et al. (2013) discovered distinct associations between more than 500K DHSs and promoters, CpG islands, conserved elements, and transcription factor motif enrichment using correlation between matched DNase-seq and gene expression data from more than 70 cell-types. Varley et al. (2013) used a diverse collection of 82 human cell lines and tissues to correlate CpG methylation and gene expression. However, those methods were limited by the availability of genomic data - a large panel of cells, with comparable quality and resolution across all conditions is required to build an accurate model of *enhancer activity ~ gene expression*. However, the optimal number of cell types to be considered in the enhancer-gene pairs annotation is yet to be determined (Hariprakash and Ferrari 2019), and thus, the reg2gene relies on the largest collection of epigenomics or transcriptomic data in humans at time - the Roadmap datasets (Bernstein et al. 2010, Roadmap Epigenomics Consortium et al. 2015).

The main advantage of correlation approaches is that they can identify multiple targets of an enhancer and directly derive a quantitative measure of the strength of association (Hariprakash and Ferrari 2019), On the other hand, the most frequently used measures of correlation, the Pearson and Spearman correlation coefficients, can only detect linear relationships. They are not sensitive to the nonlinear regulatory relationship that tends to be very common in biology (Brunel et al. 2010) and especially prevalent among gene regulatory networks (Guo et al. 2014). However, to detect nonlinear dependence for two variables with arbitrary dimensions one can use distance correlation (Székely et al. 2007). DC has proven its power and computational effectiveness to capture both linear and non-linear relationships (Gorfine et al. 2012). In addition, distance correlation estimates are quite simple without any distribution assumption (Guo et al. 2014). Thus, along with Pearson and Spearman correlation, we additionally implemented distance correlation in the reg2gene pipeline.

However, correlation does not directly consider the fact that multiple enhancers can act on a gene in a cooperative fashion (Reuter et al. 2015). To account for that, we implemented the elastic net - a regularized regression approach (Zou and Hastie 2005) - and random forest (Breiman 2001). Random forest has recently become a popular machine learning technique in bioinformatics (Zhang and Ma 2012), due to its ability to give a measure of feature importance (Bylander 2002), run efficiently on large datasets without over-fitting, and its inherently non-parametric structure (Friedman et al. 2001). It has been frequently used for variable selection, prediction modelling, pathway analysis, genetic association, epistasis detection, etc. (Diaz-Uriarte and de Andrés 2005, Chen and Ishwaran 2012). For example, in the case of enhancer predictions, Rajagopal et al. (2013) identified sets of chromatin marks that appeared to be the most informative and robust across cell-types and replicates using random forest. We mainly implemented the random forest algorithm due to the "grouping property" of RF trees (Ishwaran et al. 2010) that enabled us to adeptly deal with correlation and interaction between enhancer activity and gene expression.

Likewise, elastic net has been used to solve biological problems that require modelling of high-dimensional data with few examples and strongly correlated predictors (Zou and Hastie 2005) - such as epigenomic data. For example, EMERGE used elastic net algorithm to predict genomic regulatory elements (enhancers) from multiple genomic signatures (van Duijvenboden et al. 2016), whereas results of genome-wide association studies were analyzed with elastic net to prioritize risk-associated SNPs (Wu et al. 2009) or identify gene-gene interactions (Park and Hastie 2008). Since, it simultaneously assesses the strength of an impact of enhancers (assess the errors terms in predicting TSS activity based on the activity of all candidate enhancers) on their target genes and identifies significant models we implemented elastic net algorithm to account for the cooperative activity mode of multiple enhancers.

To assess the statistical significance for each model, reg2gene calculated the binary score that defined whether the association between enhancers and target genes was statistically significant or not. Thus, similar to GeneHancer (Fishilevich et al. 2017), we allowed more flexible prioritization of EGAs by adjusting a single threshold on the score. Then, binary scores were integrated using the majority voting approach. With this additional data integration step, we aimed to overcome the non-specificity of individual chromatin marks to predict enhancer regions in the genome (as suggested in Ernst et al. 2011).

Similar to reg2gene, JEME combined regression-based (lasso and elastic net) and supervised ML modelling (random forest, Cao et al. 2017), however, none of the previous methods reached the scope of reg2gene modelling. Especially since, we modelled each putative enhancer-gene interaction a total of twenty times: we quantified four enhancer activity scores (H3K4me1, H3K27ac, DNAme and DHS) and performed modelling using five algorithms. Since a consensus about which epigenomic or transcriptomic data assess enhancer activity the best does not exist (Shlyueva et al. 2014), we selected chromatin marks that were previously suggested to be more robust (although not perfect) predictors of enhancer activity than other marks (Whalen et al. 2016) and reported within the Roadmap datasets (Roadmap Epigenomics Consortium et al. 2015). Although, we did not directly model other predictors of enhancer activity, such as eRNAs - bi-directional non-coding RNAs produced at the location of enhancers which expression level correlates well with the functional activity of the enhancer (Mikhaylichenko et al. 2018), other methods, which predictions we integrated as the last step of our analysis, did.

Specifically, we voted our predictions together with predictions from other four computational method that map *cis*-regualtory interactions: JEME - joint effect of multiple enhancers (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS - FDR-corrected OLS with Cross-validation and Shrinkage (Hait et al. 2018) aiming to improve the reproducibility of our results and account for the fact that we might have had missed some of the associations that were previously identified across multiple publications (that used different epigenomes as an input than we did). We again relied on the approach from Yang et al. 2010 and treated EGAs as multiple classifiers that were combined by majority voting to narrow down the hypothesis space and to report more accurate models than individual predictions. Thereby, by integrating reg2gene predictions together with GeneHancer or EnhancerAtlas, we indirectly used information from the FANTOM5 (Andersson et al. 2014), and GTEx consortium (GTEx Consortium et al. 2017), GRO-Seq, etc.

All reg2gene functions were gathered in the reg2gene R package that, additionally to the enhancer-gene associations modelling, enables solving of similar biological problems such as modelling across cell types, tissues or time course series (for example, the time course of embryonic development in zebrafish, etc. (White et al. 2017).

.....

Prior to implementing the reg2gene algorithm, we made three important choices: a) identify the relevant gene annotations, b) define enhancer regions for which we will quantify activity signal and run modelling, c) define chromatin marks which signal will be used to quantify enhancer activity. First, for each gene, we identified exon locations in the GENCODE database (Harrow et al. 2012), one of the most comprehensive databases of protein-coding genes and their features. Second, since the initial choice of epigenomic marks used to map enhancers has a large impact on the final number and characteristics of defined enhancers (Zentner and Scacheri 2012), and consequently, on the number and properties of defined EGAs (Hariprakash and Ferrari 2019), we improved the robustness of our predictions by identifying a consensus of enhancers. We first used the Roadmap datasets as the largest collection of epigenomics or transcriptomic data in humans at time to identify enhancer regions predicted across its 127 cell types/tissues (Roadmap Epigenomics Consortium et al. 2015). Similar to JEME that used the same dataset to predict enhancers, we relied on ChromHMM predictions of chromatin states (Ernst and Kellis 2012) and identified 184,005 genomic locations that were commonly predicted as enhancer-like across many different cell types. We next considered sets of enhancers from EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), and FOCS (Hait et al. 2018), and our *in-house* definition to represent different classifiers (that define individual hypothesis spaces) and their integration was proposed to narrow the hypothesis space and return more accurate predictions (Yang et al. 2010). In other words, we defined 286,723 "consensus" enhancers as regions in the genome that were covered by at least three enhancer definitions. This number represents a relatively small subset of initially estimated 1 million enhancer regions (Heintzman et al. 2009, ENCODE Project Consortium 2012) or 2.9 million DHSs (DNase I hypersensitive sites; Thurman et al. 2012), but nevertheless outperforms 43,011 enhancer candidates from the FANTOM5 consortium (Andersson et al. 2014). On the other hand, as a consensus of enhancer definitions, this number is smaller than the number of individual JEME enhancers (~300K), but larger than the number of GeneHancer and EnhancerAtlas enhancers (~200K). However, an exhaustive reference list of enhancers is missing and their true number is still unknown (Shlyueva et al. 2014, Hariprakash and Ferrari 2019).

Although regulatory elements may have both enhancer and promoter functions (Andersson et al. 2015; Andersson and Sandelin 2020), the percentage of promotes that have a strong enhancer activity in the genome was shown to not be large (3% of tested gene promoters in human K562 cells, Dao et al. 2017), and we hypothesized that "better" enhancer predictions should have a smaller overlap with promoters. As compared to previously defined enhancers, our enhancer

regions rarely overlapped with promoter regions (>5% of *in-house* enhancers as compared to 20% GeneHancer enhancers). As expected, ChromHMM-predicted enhancer-like states (6_EnhG and 7_Enh) showed a greater per-base overlap with in-house enhances (45-51%), as compared to 0% of covered heterochromatin regions or 2% of quies.

.....

Prior to reg2gene modelling, we paired the "consensus" enhancers with genes to define 8.9 million enhancer - gene pairs of which ~1M was statistically significant and further voted together with JEME, GeneHancer, EnhancerAtlas and FOCS EGAs. We defined two datasets of voted EGAs: ~230K EGAs in the *flexibleC* set (EGAs voted by two methods), and ~60K *stringentC* EGAs (EGAs voted by three methods). Datasets were built on top of ~110K and ~45K enhancer regions that altogether indicated that predictions are genome-wide. As compared to the previous computationally assessed EGAs, we report the most stringent dataset in every aspect: number of genes, enhancers, EGAs, an average of genes reported per enhancer region or vice versa enhancers per gene. For example, we calculated that EnhanceAtlas enhancers were paired with a median of 24 genes. This large number of interactions could be a consequence of the fact that EnhancerAtlas-defined enhancers were on average longer than enhancers defined in other publications (up to 3Mb), but as well we might have inflated the statistics with our protocol that pooled (and reduced) all enhancer-gene associations across 105 analyzed cell types. Indeed, Gao et al. 2016 reported that one EnhancerAtlas enhancer was associated with 2.4 target genes, and each gene was associated with 4.1 enhancers when an average was calculated in individual cell types. Likewise, we calculated that GeneHancer enhancers participate, on average, in 14 interactions per gene (median), but Fishilevich et al. 2017 reported 1.44 genes per enhancer and 7.47 enhancers per gene. In FOCS, each promoter was reported to be linked to 2.4 enhancers or specifically, in optimally reduced models, each promoter was linked, on average, to 3.2, 2.8, and 3.6 enhancers in the Roadmap, FANTOM5, and GRO-seq datasets (Cao et al. 2017). Those numbers are similar to a mean of 1.94 genes per enhancer and a median of 9 for enhancers per gene for *flexibleC* and mean of 1,25 genes per enhancer for *stringentC* models. Nevertheless, as we still do not have a precise knowledge about the total number of enhancer regions in the human genome (Shlyueva et al. 2014), nor the number of their interactions (Hariprakash and Ferrari 2019). Thus, we do not know which reported statistics represent the correct population average. However, since *in vivo* testing of the reporter or *in vivo* editing of the enhancer in transgenic animals has been considered to be a definitive proof of enhancers (Visel et al. 2007; Catarino and

Stark 2018), we considered the estimation of Fulco et al. (2019) - individual enhancers regulate up to 5 genes, whereas individual genes can be regulated by up to 12 distal elements - to be the "most" precise estimations at time.

Due to the fact that boundaries between topologically associating domains have specific insulating properties, TADs are generally defined as regions in the genome characterized by a high level of chromatin interactions occurring within them, many of which can be enhancer-gene interactions (Nora et al. 2012; Lupiáñez et al. 2015). Thus, we hypothesized that the higher percentage of EGAs within topologically associating domains denotes better predictions of regulatory associations and we calculated the percentage of colocalizing events. Since TADs show a general conservation in relative position across different cell types and/or organisms (Dixon et al. 2012; Vietri Rudan et al. 2015), we expected to observe that the number of interactions reported in different cell types should not vary much, and encouraged us to calculate an average colocalization percentage across different cell types. Except in the case of *inhouseM* EGAs, we confirmed our expectation that EGAs are mostly colocalized in the same TAD region, and showed that, on average, more than 80% of *stringentC* and FOCS predicted EGAs were located in the same TAD region in different cell types.

4.5. Conclusions

We developed and implemented the reg2gene algorithm - a computational method that maps cis-regulatory interactions in the human genome. Although reg2gene itself performs computational modelling, we integrated its results (predicted enhancer-gene associations) together with four other sets of enhancer-gene associations: EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), and FOCS (Hait et al. 2018). With this approach, we aimed to produce results that are more robust. To analyze whether or not that holds truth, in the following chapter, I set off to systematically benchmark sets of enhancer-gene associations.

5

Results III: Benchmarking of reported enhancer-gene associations

Preface

In this section, I benchmark results of five computational methods that map cis-regulatory interactions in the genome-wide manner. I defined the scope of this analysis mainly together with Dr. Vedran Franke, and supported by Dr. Altuna Akalin.

Abstract

Multiple approaches have been utilized to study enhancer-mediated long-range gene regulation in a genome-wide manner: predictions using information from the eQTL studies (GTEx Consortium et al. 2017), (3C)-derived techniques (Lieberman-Aiden et al. 2009; Fullwood et al. 2009), reporter assays or cellular screens (Arnold et al. 2013; Gasperini et al. 2019); and their results have been frequently used to benchmark computational models of gene expression ~ enhancer activity.

In this chapter, I benchmark enhancer-gene associations assessed by reg2gene and other computational methods: JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS (Hait et al. 2018) using eQTLs from two studies (GTEx Consortium et al. 2017, Westra et al. 2013), and chromatin interactions from two data sources (Xie et al. 2016, Javierre et al. 2016).

First, I show that none of the used benchmark datasets can be considered to represent a "golden" standard that can ultimately exhibit differences in the performance between computationally methods that assess enhancer-gene associations. Thus, we additionally defined a set of positive and negative EGAs by combining results of cellular screens (Gasperini et al. 2019) and defining an "in-house" EGA negatives using an extensive data integration approach (Ernst and Kellis 2012). With such defined benchmark dataset, we showed that stringentC models indeed have the highest positive predictive value (PPV) across all seven sets of analyzed EGAs.

5.1. Introduction

Multiple computational methods have been developed to assess *cis*-regulatory interactions in the human genome. Previously, we systematically analyzed several of them: JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), and FOCS (Hait et al. 2018) and showed that their protocols and predictions vary tremendously: especially in the location, number and properties of defined enhancer regions and enhancer-gene associations. Next, we developed a novel method - the reg2gene algorithm - that, in addition to the computational modelling of *gene expression ~ enhancer activity*, integrates its results (enhancer-gene associations) with results of the aforementioned methods. Finally, here, we compare their results and a potential to be benchmarked (or benchmark).

Up until today, multiple test datasets have been frequently used to benchmark computationally predicted EGAs (Gao et al. 2016, Cao et al. 2017, Hait et al. 2018). They can be broadly categorized into predictions of EGAs using information from: the eQTL studies (GTEx Consortium et al. 2017), chromosome conformation capture (3C)-derived techniques (Lieberman-Aiden et al. 2009; Fullwood et al. 2009), or direct functional confirmation of enhancer activity by reporter assays (Arnold et al. 2013) and cellular screens (Arnold et al. 2013). In other words, eQTL-eGene associations or reported chromatin interactions have been considered to represent a good proxy (but not a direct proof!) of enhancer-mediated regulation of gene expression; especially since each of the aforementioned high-throughput technologies/datasets is characterized by specific or common limitations marked by noisy data and high number of false negatives or positives (Hariprakash and Ferrari 2019).

This opens a question, which, if any, of the aforementioned datasets should be used for benchmarking and what would be results. To answer those questions, in this chapter, I benchmark seven sets of enhancer-gene associations using six different experimental datasets and systematically analyzed results.

5.2. Methods

5.2.1. The intuition behind the benchmarking protocol

We used two eQTL studies as benchmark datasets (GTEx Consortium et al., 2017, Westra et al., 2013). To benchmark with eQTL-eGene associations, we firstly identified eQTLs that overlap with enhancers and then, for each overlapping eQTL-enhancer pair, we compared whether the eQTL-associated eGene (gene associated with the given eQTL) equals enhancer-annotated gene (**Figure 5.1.**). If yes, that pair was considered to be part of the benchmark dataset.

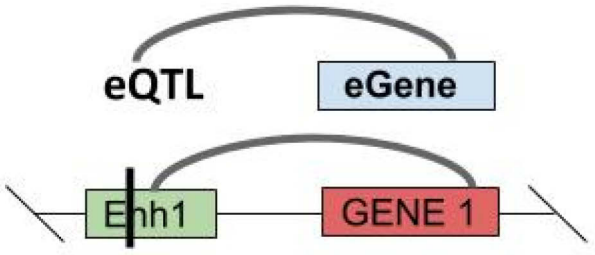


Figure 5.1. Schematic representation of benchmarking EGAs with pairs of eQTL-eGene associations. First, we would identify eQTLs that overlap with enhancers and then, for each overlapping eQTL-enhancer pair, we compared whether the eQTL-associated eGene (gene associated with the given eQTL) equals enhancer-annotated gene. If yes, that pair was considered to be benchmarked.

We benchmarked sets of EGAs with chromatin interactions from the PC-HiC experiment (Javierre et al., 2016) and CCSI database (Xie et al., 2016). Since, both anchors of chromatin interactions (locations of the interacting pair are commonly referred to as an anchor one and anchor two) could potentially overlap the TSS or enhancer location, we tested an overlap twice, in both orientations. We first identified an overlap between the enhancer regions and anchor one of the interacting pair, and confirmed it by locating the TSS within the anchor two region. Then, we tested if there was an overlap between the enhancer region and anchor two of the interacting pair, and anchor one and gene location. If we identified an overlap in any of these two overlapping steps, we considered the analyzed enhancer-gene association to be benchmarked with the given benchmark dataset (**Figure 5.2.**).

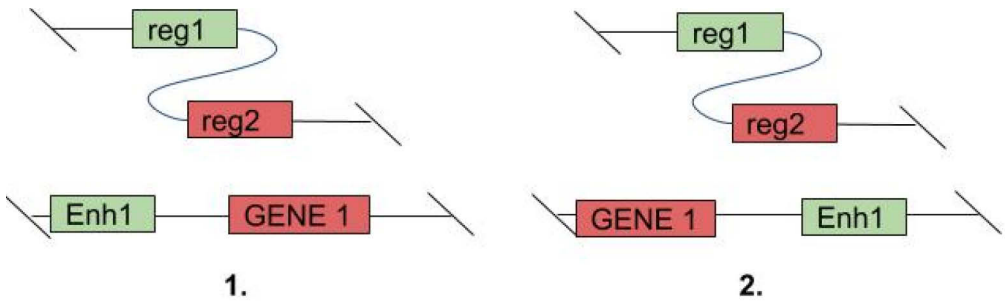


Figure 5.2. Schematic representation of benchmarking EGAs with pairs of chromatin interactions. We tested an overlap between both anchors of chromatin interactions and the TSS-enhancer pair twice; anchors in both orientations are compared to the tested enhancer-gene pair.

Importantly, prior to benchmarking, we accounted for the fact that many EGAs do not have a potential to be benchmarked at all - they are located outside of any region that is present in the benchmarking dataset.

5.2.2. An overview of used datasets

I used multiple datasets to benchmark assessed interactions (**Table 5.1.**): eQTL studies (Westra et al., 2013, GTEx Consortium et al., 2017), the 3C-technology based studies (the PC-HiC study by Javierre et al., 2016) and the CCSI database (Xie et al., 2016), result of a single-cell cellular screen (Gasperini et al., 2019).

Table 5.1. Characteristics and statistics behind datasets used in this chapter

Data source	Technology/Data type	Processing algorithm	Used as/for:	N of EGAs/interactions
GTEx database (GTEx Consortium et al., 2017)	Microarrays/eQTLs	NA	Benchmarking	388,160
(Westra et al., 2013)	Microarrays/eQTLs	NA	Benchmarking	672,717
(Javierre et al., 2016)	PC-HiC	NA	Benchmarking	728,838
CCSI database (Xie et al., 2016)	3C, 4C, 5C, ChIA-PET and Hi-C	NA	Benchmarking	1 587 002
(Gasperini et al., 2019)	dCas9-KRAB	NA	Benchmarking	449
Lieberman-Aiden et al. 2009, Rao et al. 2014, Dixon et al. 2015, Leung et al. 2015, Schmitt et al. 2016, ENCODE Project Consortium 2012	Hi-C	multiple	Assessing overlaps with TADs	NA

5.3. Results

5.3.1. Validation of our benchmarking procedure

Authors of the previous computational EG association methods used eQTLs and chromatin interactions as a "golden benchmark" datasets to benchmark their predictions and assess the performance of their algorithms (Cao et al. 2017, Hait et al. 2018). To get a better understanding of the benchmarking problematics, we, as well, benchmarked our predictions using those two datasets.

However, contrary to the previous publications that mostly benchmarked using one dataset, we set off to systematically benchmark *cis*-regulatory interactions using **two** eQTL studies (GTEx Consortium et al. 2017, Westra et al. 2013), and **two** sources of chromatin interactions (Xie et al. 2016, Javierre et al. 2016) which largely varied in their sizes (**Supplementary Figure 2.**). We specifically opted to use benchmark datasets that were reported to provide a good proxy information of enhancer-mediated gene regulation (Hariprakash and Ferrari 2019) and/or that were utilized by the previous publications to benchmark their results (Cao et al. 2017, Hait et al. 2018).

As the primary source of eQTLs, we used the GTEx database - a largest repository of tissue-specific eQTLs (GTEx Consortium et al. 2017). We additionally included the second source of eQTLs as a control - Westra et al. 2013 - that we specifically selected because it reported eQTLs across a large number of cell types, as well as *trans* eQTLs. Contrary to *cis* eQTLs, *trans* eQTLs alter the structure, function or expression of a diffusible factor (Ronald et al. 2005), and thus act mostly indirect, non-allele specific and in long ranges. The last property of *trans* eQTLs was the most attractive to us, especially since we were primarily focused on identifying long-range interactions and expected to observe many benchmarked *trans* eQTLs.

Since the GTEx eQTLs were previously used as the benchmark datasets in FOCS (Hait et al. 2018) and JEME (Cao et al. 2017), we used this information to control for our benchmarking procedure. Around 170,000 (37%) and 90,000 (38%) of FOCS enhancer-promoter links were reported to be supported by ChIA-PET from the CCSI database (Xie et al. 2016) and GTEx eQTLs (GTEx Consortium et al. 2017), respectively. However, using the CCSI database we managed to confirm around

60,000 FOCS enhancer-promoter links, whereas the GTEx eQTLs confirmed only around 2,500 FOCS E-P links. After we repeated the benchmarking, but with extended enhancers (+/-500bp) as they did in the original publication, we managed to obtain a 10x increase in the number of overlaps, and with this, we roughly confirmed the reported percentages, and verified our benchmarking procedure.

In the case of chromatin interactions, we corroborated previous results by FOCS (Hait et al. 2018) and confirmed that ~60% of their EGAs is confirmed by chromatin interactions from the CCSI database (although, in their case they only tested ChIA-PET experiments, Xie et al. 2016).

As a second benchmark dataset, we used two sources of 3C-related experiments: the promoter-capture HiC (PC-HiC) experiment (Javierre et al. 2016) and the CCSI database (Xie et al. 2016). We relied on the notion that the (3C)-derived techniques measure chromatin contact frequencies, which correlate well with functional studies of regulatory elements (Pombo and Dillon 2015), and thus can be used as a good proxy information of enhancer-mediated gene regulation. The CCSI gathers information from all previously conducted 3C, 4C, 5C, ChIA-PET and Hi-C experiments but is not an up-to-date database. On the other hand, as we were specifically interested in the analysis of gene expression regulation, we used a genome-wide dataset that is strongly enriched for promoter interactions - PC HiC (Javierre et al. 2016).

5.3.2. The eQTL studies and chromatin interactions identified by (3C)-derived high-throughput technologies suffer from low reproducibility

To test the actual benchmarking potential of four benchmarking datasets and whether the two eQTLs studies reported more similar results as compared to the chromatin interaction studies (and vice-versa if the reported chromatin interactions are more similar to each other than to the eQTL-eGene associations), we analyzed the characteristics and the performance of benchmark datasets to predict each other's results (**Figure 5.3.**).

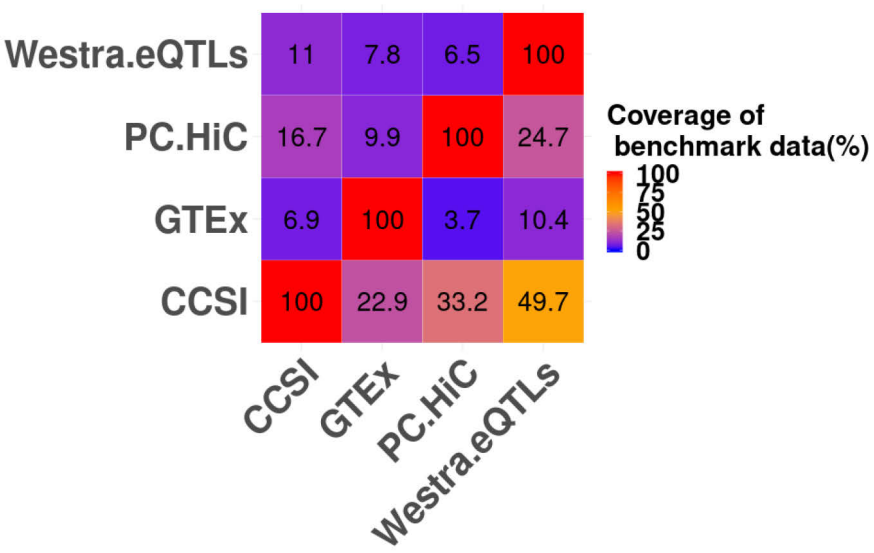


Figure 5.3. Heatmap of the benchmarking results for two eQTL-eGene association studies and chromatin interactions reported in (Javierre et al., 2016) and the CCSI database (Xie et al., 2016). Overlap percentage was reported for studies in columns. For example, 10.4% of the eQTLs from the GTEx database were confirmed by eQTLs reported in Westra et al., (2013)

We identified that one third of the PC-HiC chromatin interactions (33.2%) could be benchmarked by the CCSI-reported interactions, whereas 16.7% of the CCSI database is supported by the PC-HiC-identified chromatin interactions from Javierre et al., 2016. Westra eQTLs overlapped 10.4% of the GTEx eQTL-eGene associations and vice-versa 7.8% GTEx eQTLs associations was covered by the Westra eQTL-gene links. Less than 4% and 7% of PC-HiC interactions were supported by the GTEx and Westra eQTL-eGene associations, respectively, as compared to 7% and 11% of CCSI covered interactions. The GTEx eQTLs achieved the greatest overlap with the CCSI database (23%), but only 10% with another source of chromatin interactions - PC-HiC. On the other hand, 50% of the Westra eQTLs overlap with the CCSI chromatin interactions and 25% of PC-HiC interactions. Altogether, this indicated that the analyzed benchmarking datasets suffer from poor reproducibility. With this in mind, we performed a thorough benchmarking analysis.

5.3.3. Different sets of EGAs are diversely covered by eQTL-eGene pairs (eQTLs) and chromatin interactions

As expected, EGAs from different datasets were unequally covered by both sources of interactions: a total of 7.5%, 9.1%, 11.5% and 18.8% of EnhancerAtlas, FOCS, JEME and GeneHancer enhancer-gene associations, respectively, were supported by the GTEx eQTLs (**Figure 5.4., Supplementary Figures 3. and 4.**). However, only 1.4% of the *inhouseM* was confirmed by eQTLs but this percentage increased to 7.3% for *stringentC* models. In the case of Westra et al., (2013) eQTLS, we observed an increase in the overlap between EGAs and eQTL-eGenes for each defined set of EGAs; 18%, 16%, 23.8% and 25.4% for EnhancerAtlas, FOCS, JEME and GeneHancer, respectively. Although only 3.5% of the *inhouseM* EGAs overlapped Westra eQTLs, this percentage increased to 9.5% and 18.2% for *flexibleC* and *stringentC* EGAs.

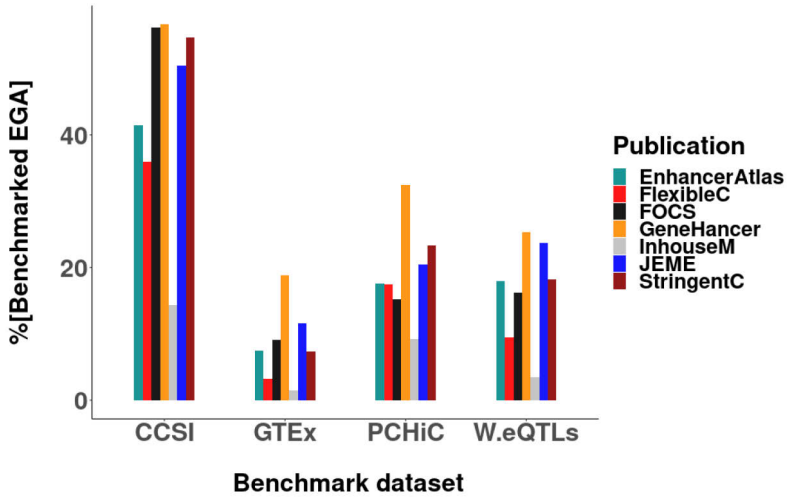


Figure 5.4. Histogram of the percentage of overlap between publication-specific enhancer-gene associations (EGAs) and four different benchmark datasets: two eQTL studies (Westra et al., 2013), GTEx database, (GTEx Consortium et al., 2017) and two sources of HiC-related experiments (Javierre et al., 2016), CCSI database (Xie et al., 2016). Percentage overlap was calculated for each of the seven sets of EGAs. Benchmarked interactions are those in which one location (anchor one) of the interaction pair overlaps an enhancer region, and the other location (anchor two) overlaps extended gene's TSS, or vise-versa anchor two overlaps enhancer, whereas anchor one overlaps with promoter. On the other hand, if an eQTL-associated gene is equal to the gene associated with the overlapping enhancer, such eQTL-eGene was considered to be benchmarked.

On the other hand, 41%, 55%, 50% and 56% of EnhancerAtlas, FOCS, JEME and GeneHancer enhancer-gene associations, respectively, were supported by the CCSI database, whereas 17.6%, 15%, 20.5% and 32.5% of EGAs were supported by the PC-HiC interactions. In the case of in-house models, 14.3% and 9.2% of the *inhouseM* was confirmed by CCSI and PC-HiC interactions. This percentage increased to 36% and 55% for the *flexibleC* and *stringentC* models and the CCSI

reported interactions; 23.4% *stringentC* interactions were reported by PC-HiC. Importantly, all interactions that do not overlap with interactions reported in the benchmark dataset were filtered out prior to this analysis.

Next, we changed the perspective and analyzed the modelling potential of computational methods to predict eQTLs and chromatin interactions.

In general, Westra et al., 2013 showed a greater coverage by computational predicted EGAs than other benchmark datasets (**Figure 5.5., Supplementary Figures 3. and 4.**). Consistent with most of other benchmark datasets the biggest percentage of overlaps was identified if EnhancerAtlas EGAs were used as a benchmark dataset, followed by GeneHancer and JEME.

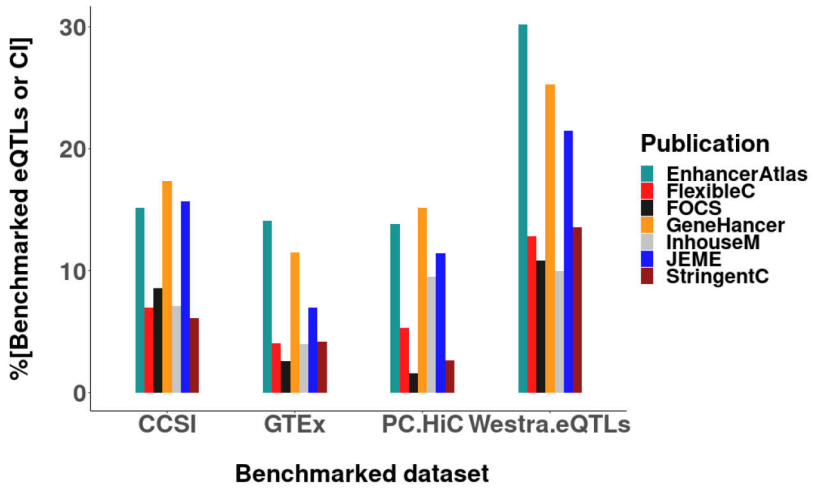


Figure 5.5. Histogram of the percentage of overlap between four different benchmark datasets: two eQTL studies (Westra et al., 2013), GTEx database, (GTEx Consortium et al., 2017) and two sources of HiC-related experiments (Javierre et al., 2016), CCSI database, (Xie et al., 2016) and publication-specific enhancer-gene associations (EGAs). Benchmarked interactions are those in which one location (anchor one) of the interaction pair overlaps an enhancer region, and the other location (anchor two) overlaps extended gene's TSS, or vise-versa anchor two overlaps enhancer, whereas anchor one overlaps with promoter. On the other hand, if an eQTL-associated gene is equal to the gene associated with the overlapping enhancer, such eQTL-eGene was considered to be benchmarked. On the y-axis is the percentage of benchmarked eQTLs or chromatin interactions identified by overlapping them with individual sets of enhancer-gene associations (EGAs).

Specifically, GTEx eQTLs showed the highest coverage with EnhancerAtlas (14%), but only 2,6% of FOCS, and 4% of *stringentC* supported eQTL-eGene GTEx associations. Compared the GTEx eQTLs, Westra et al., 2013 eQTLs showed higher coverage with all other methods to predict EGAs; 30% of eQTL-eGene associations were covered with EnhancerAtlas; 25% with GeneHancer and 21% with JEME. The most obvious difference in the prediction success was observed for EnhancerAtlas

and eQTL studies: twice more Westra et al., 2013 eQTL-eGene pairs was predicted by the EnhancerAtlas EGAs than the GTEx eQTLs. Equal discrepancy in predictions was observed for other methods as well.

On the other hand, 17.3% of chromatin interactions reported in the CCSI database was confirmed by GeneHancer EGAs, which was followed by 15.7% interactions confirmed by JEME and 15.2% by EnhancerAtlas as compared to 6-7% overlap with in-house models. Chromatin interactions observed by Javierre et al., 2016 had even lower overlap percentages: 15% of them were confirmed by GeneHancer, but only 2.7% by *stringentC*.

5.3.4. Benchmarking with "*in-house*" defined set of negative EGAs and results of cellular screens revealed that *stringentC* models have the highest PPV

We next performed the second benchmarking step, in which we used a predefined set of *in-house* EGAs negatives and include *cis* enhancer-gene interactions from cellular screening as a benchmark set (Gasperini et al., 2019).

We hypothesized that *cis* enhancer-gene interactions from Gasperini et al. (2019) correspond to a set of positive interactions, that if predicted by computational methods, represent true positives (TP), and if missed, false negatives (FN). We specifically opted for interactions reported in Gasperini et al., 2019, because they managed to capture perturbations of gene expression globally. By using single-cell RNA sequencing, and a unique combination of perturbations introduced into each individual cells with gRNAs at a high MOI (multiplicity of infection), they overcome limitations of the previous methods that either tested a single gene per experiment, or had low power due to the low multiplicity of lentivirus infection (Canver et al. 2015; Wakabayashi et al. 2016; Diao et al. 2017). We opted for cellular screens since it has been considered that only *in vivo* testing of the reporter or *in vivo* editing of the enhancer in transgenic animals represent a definitive proof of enhancers and their activity (Visel et al. 2007; Catarino and Stark 2018). However, until recently (Fulco et al. 2019, Gasperini et al. 2019), only low-throughput assays were available to provide a direct functional confirmation or quantitative readout of enhancer activity (Arnold et al. 2013, Kvon 2015).

However, to be able to fully assess the performance of EGAs methods, we needed to assess other two elements of the confusion matrix: false positives (FP) and true negatives (TN) and thus, we developed an approach that searched for "negatives'' - enhancer-gene pairs for which we expected they would not be statistically associated. To achieve that we proposed an approach that is based on data integration of Roadmap epigenomes (Roadmap Epigenomics Consortium et al. 2015) and across-cell-type ChromHMM-based chromatin states (Ernst and Kellis 2012). We used ChromHMM-predicted chromatin states (Ernst and Kellis 2012, Roadmap Epigenomics Consortium et al. 2015) because they were identified by computational modelling and data integration of multiple chromatin marks, and thus, they should overcome non-specificity of individual chromatin marks and inherent weaknesses of different high-throughput datasets (Yang et al. 2010, Ernst et al. 2011). This idea was based on the assumption that *enhancer activity ~ gene expression* modelling should not be successful for enhancers that were predicted by ChromHMM as Polycomb repressed, heterochromatin or quiescent chromatin state across all 127 cell types and consequently, EGAs for "inactive" enhancers should not be predicted as EGAs.

In summary, we first screened the 15-states ChromHMM predictions across all 127 Roadmap cell types with the location of enhancers and identified whether defined enhancer overlaps with any of the three "inactive" ChromHMM-predicted chromatin states: Polycomb-repressed, heterochromatin or quiescent. If we identified enhancer that was across all 127 cell types "inactive" (100% per base overlap with repressed, heterochromatin or quiescent ChromHMM-predicted chromatin states, **Figure 5.6.**), then we considered such enhancer region to be a false positive - incorrectly identified as an enhancer. Consequently, we consider each reported enhancer-gene pair, which enhancer was identified to be "inactive" across all cell lines to be *false positive* EGA.

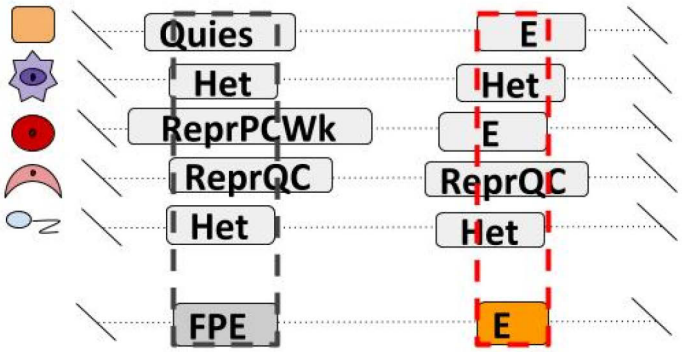


Figure 5.6. Schematic representation of defining false positive (FP) enhancers in the genome (FPE). If across all 127 cell types an enhancer was found to be inactive (100% per base overlap with repressed, heterochromatin or quiescent ChromHMM-predicted chromatin states), we considered such enhancer region to be a false positive enhancer (FPE).

We identified that, out of 321 high-confidence *cis* enhancer-gene interactions reported in Gasperini et al., 2019, 291 indeed overlapped GeneHancer and were considered true positives (TPs). A total of 180 (Gasperini et al., 2019) enhancer-gene interactions was benchmarked by EnhancerAtlas predictions, out of 210 possible. In addition, 91/114 TP was predicted by FOCS, 129/169 by JEME. In case of our inhouse models, a total of 45, 61, 67 of *stringentC*, *flexibleC* and *inhouseM* EGAs overlapped (Gasperini et al., 2019) defined positives, and thus, were identified as TP from this subset (**Figure 5.7.**, **Supplementary Table 2**).

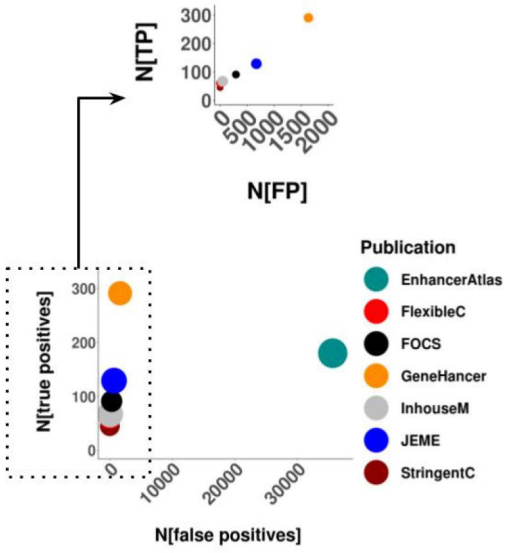


Figure 5.7. Results of the overlapping procedure for the inhouse sets of expected negative and positive enhancer-gene associations. Seven sets of computationally predicted EGAs were tested for an overlap. Number of identified false positives (FP) is reported on the x-axis, whereas true positives (TP) are reported on the y-axis. TP correspond to *cis* enhancer-gene interactions reported in (Gasperini et al., 2019) that overlapped computationally predicted EGAs. FP are computationally predicted EGAs whose enhancers were predicted to be "inactive" by the ChromHMM algorithm across all 127 cell types reported in the Roadmap dataset (details in Methods).

In the case of negatives, we previously identified that 272 JEME, 274 FOCS, 1121 GeneHancer, and 13,672 EnhancerAtlas defined enhancers overlap “inactive enhancer” regions present across all 127 epigenomes (**Supplementary Table 1.**). Nevertheless, we identified that those enhancers participated in 35,618 EnhancerAtlas EGAs or 1,635 GeneHancer EGAs. On the other hand, we identified 126 *consensusE* that were predicted by ChromHMM as repressed, heterochromatin or quiescent across all Roadmap cell types. Paired with genes, “inactive” *consensusE* enhancers built 3,147 enhancer-gene pairs, each of which was modelled the reg2gene algorithm. However, only 12 enhancers were reported to have significant association in the *inhouseM* EGAs, whereas 3 enhancers had significant EGAs in the *flexibleC* models. None of the enhancers from the stringentC EGAs were identified to be “inactive”.

Overall, this enabled us to fully assess the performance of computational methods by assessing all elements of the confusion matrix: false positives (FP), false negatives (FN), true negatives (TN) and true positives (TP). For example, by using defined negatives (expected “inactive” enhancer-gene pairs) we could assess true negatives (TN) and false positives (FP), whereas *cis*-interactions from Gasperini et al. (2019) allowed us to identify true positives (TP) and false negatives (FN). With quantified true and false positives, we could calculate positive predictive value (PPV) as the number of true positives (TP) over all positives (PPV=TP/(TP+FP). The highest PPV value was observed for *stringentC* models (100%), followed by *flexibleC* models (93.4%). On the other hand, although GeneHancer reported the highest number of true positives (N=292) its PPV was 15.1%.

5.4. Discussion

Multiple computational methods have been developed to map *cis*-regulatory interactions in the human genome. In this chapter, I assessed the potential of our *in-house* method - reg2gene - and several other computational algorithms: JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), FOCS (Hait et al. 2018) to predict eQTLs and chromatin interactions, or have their results benchmarked. Since reg2gene provided us with three sets of predictions: *inhouseM* EGAs (enhancer-gene associations that were a direct result of integration of twenty different models of *gene expression~enhancer activity*) and *stringentC* and *flexibleC* EGAs (two consensus sets of EGAs that were voted based on EGAs from *inhouseM,* JEME, GeneHancer, EnhancerAtlas, and FOCS) we separately benchmarked all three of them.

As benchmark datasets, we used two sources of eQTLs (Westra et al. 2013, GTEx Consortium et al. 2017) and chromatin interactions (Javierre et al. 2016, Xie et al. 2016), results of cellular screens (Gasperini et al. 2019) and the ChromHMM-predicted repressed regions in the genome (Ernst and Kellis 2012) and performed a total of thirty-five pairwise benchmarking analyses.

First, we verified the accuracy of our benchmarking protocol by comparing our benchmarking results for GTEx eQTLs with the results reported in FOCS (Hait et al. 2018) and JEME (Cao et al. 2017). Specifically, Hait et al. (2018) reported that FOCS outperforms extant methods in terms of concordance with enhancer-promoter interactions identified by ChIA-PET from the CCSI database (Xie et al. 2016), HiChIP (Weintraub et al. 2017), and eQTL data (GTEx Consortium et al. 2017). We revealed a much lower percentage of overlaps between eQTLs and enhancer-gene associations as compared to the reported ones in FOCS, however, after we repeated the benchmarking procedure, but with extended enhancer regions (as they did in the original publication), we obtained the comparable results, thus verifying our protocol. On the other hand, Cao et al. (2017) compared JEME with other state-of-the-art methods for predicting enhancer targets on the basis of overlap with eQTLs (Ernst and Kellis 2012, He et al. 2014, Corradin et al. 2014, Roy et al. 2015, Whalen et al. 2016) and reported that JEME was the most accurate of all the methods in across-sample tests, whereas in cross-validation tests, JEME was slightly less accurate than TargetFinder (Whalen et al. 2016). However, Cao et al. (2017) reported AUPR, or area under the curve, as a performance statistics that requires quantification of all elements of the confusion matrix: true positives, false negatives, true negatives and false positives. To count them, they hypothesized that eQTLs are the “golden” standard method for assessing enhancer-gene associations and

benchmarking. However, this is not necessarily the case (and will be further elaborated later in this discussion) especially since eQTL microarrays and RNA-seq technologies suffer from many technical limitations (Ellis et al. 2013). In addition, with the second test of our benchmarking procedure, e.g. test of assessed and reported concordance between FOCS EGAs and interactions reported in the CCSI database (Xie et al. 2016), we corroborated the reported percentage of overlaps (although, in their case they only tested a subset of interactions from ChIA-PET experiments).

Since we were aware that our analysis suffered from many biases and confounding factors, we expect to observe that the final result of multiple benchmarking procedures would look different when different high-throughput datasets would be used as a "golden" benchmark. First, eQTLs and chromatin interactions are not a direct proof of enhancer-mediated gene expression regulation and they suffer from technical and biological limitations (Hariprakash and Ferrari 2019). For example, inherent biases in restriction enzyme cutting efficiency, ligation frequency, averaging of the chromatin states across the population of cells affect specificity of assessed interactions (O'Sullivan et al. 2013). Likewise, since enhancer-promoter loops may be detectable in cell types where the target gene is not active (Rao et al. 2014), or their interactions may precede the activation of target genes (Stadhouders et al. 2018), the detection of chromatin interactions with (3C)-derived methods does not unambiguously prove the presence of an active regulatory interaction in a given cell type. Ligation events detected by (3C)-derived technologies could as well reflect the higher-order nuclear organization contacts (Gavrilov et al. 2013) or random contacts between chromatin (Dekker et al. 2013), which adds to the background noise and imprecise detection of chromatin interactions (Pombo and Dillon 2015). In addition, with present technologies such as HiC, the appropriate resolution is hardly achievable for the ETG pairing tools; for example, in ChIP-seq (and consequently ChIA-PET) the ultimate resolution limit is the chromatin fragmentation size, which is usually in the order of a few hundred base pairs (Park 2009). On the other hand, eQTLs rely on microarray technologies, and as such, they are hindered by their specific technical limitations - for example, they do not interrogate the whole genome, whereas they focus on specific single nucleotide positions in the genome (Hacia et al. 1999) that are not necessarily the causal ones. "Index" SNPs are most likely only "proxy" SNPs which are in the linkage disequilibrium (LD) with the causal ones (Schork et al. 2009; Maurano et al. 2012; Tak and Farnham 2015). RNA-seq processing itself is not without its own challenges, given that mapping bias (Vijay et al. 2013, Panousis et al. 2014), coverage issues, outlier samples, batch

effects, and unknown covariates (Leek and Storey 2007) limit data integrity and eQTL reproducibility (Ellis et al. 2013).

Thus, we first cross-compared the benchmarking datasets to get a better overview of the underlying confounding factors. In agreement with what was recently reported (Forcato et al. 2017; Lajoie et al. 2015; Dali and Blanchette 2017), we detected a discordance between two sets of chromatin interactions: only one third of PC-HiC interactions was previously reported in the CCSI database. This might be a consequence of the analysis that was performed with different datasets (originated from different cell types). In other words, enhancers are not being active in cell types in which chromatin interactions were identified. However, this discrepancy more likely reflected the aforementioned limitations of (3C)-derived technologies - reported chromatin interactions do not unambiguously prove the presence of an active regulatory interaction in a given cell type (Pombo and Dillon 2015). In addition, the choice of the algorithm used for HiC analysis and TAD calling was shown to have a strong impact on the final annotations of TADs**,** especially in terms of TAD numbers and sizes (Zufferey et al. 2018; Wu et al. 2020). This variability potentially reflects an underlying hierarchical domain organization that is only partially captured by different methods and at different resolutions (Zufferey et al. 2018). More importantly, the reproducibility of Hi-C loops was shown to be low at all resolutions; even lower than the reproducibility of TAD boundaries (Forcato et al. 2017).

On the other hand, only 8-10% of eQTLs from one eQTL benchmarking dataset was present in the second one (an overlap between GTEx and Westra eQTLS). Small reproducibility of eQTLs between samples and tissues has been a long recognized problem (Dimas et al. 2009). However, previous studies disagree about the underlying replication rate of eQTLs - for example, Innocenti et al. (2011) replicated 67% of the previously identified eQTLs from primary human liver tissue, whereas two independent sets of lymphoblastoid cell lines showed ~83% overlap in reported eQTLs (Ding et al. 2010). Two sets of eQTLs that are originating from two tissues showed much lower reproducibility rate: 8.1% of the eQTLS from the prostate was confirmed by eQTLs identified in blood (Larson et al. 2015), 25% of local, and presumably, *cis*-acting eQTLs were found to be shared between blood (PBMCs) and brain (Dimas et al. 2008), and ~30% eQTLs are shared between blood and adipose (Dixon et al. 2007). Thus, a discrepancy between two sets of eQTLs can be caused by the fact that GTEx eQTLs were tested across 44 tissues, whereas in Westra et al 2013 gene expression intensities were measured exclusively in the whole blood samples. However, we did not expand our search by including all other SNPs that are in high LD with the index SNP - this

could lead to higher reproducibility rate of eQTLs. Nonetheless, such a low reproducibility of benchmark datasets opened a question, which, if any, of the aforementioned datasets should be used for benchmarking. With that in mind, we benchmarked seven sets of enhancer-gene associations, and vice-versa tested which eQTLs and chromatin interactions can be confirmed by EGAs.

.....

As expected (due to their discern properties), EGAs from all seven sets of analyzed EGAs (JEME, GeneHancer, EnhancerAtlas, FOCS, *inhouseM*, *stringentC* and *flexibleC*) were diversely covered by eQTL-eGene pairs (eQTLs) and chromatin interactions. In general, GeneHancer EGAs showed the highest percentage of coverage by both eQTLs and chromatin interactions across all benchmark datasets. This did not surprise us because Fishilevich et al. (2017) used eQTLs form the GTEx database (GTEx Consortium et al. 2017) and CHiC datasets (Mifsud et al. 2015) as an input data, e.g. one of the methods to map *cis*-regulatory interactions. Likewise, JEME EGAs had the second highest overlap percentage with eQTL datasets. Again, Cao et al. 2017 "*trained multiple enhancer–TSS pairs based on 'gold standard' answers defined by a set of validation data (ChIA-PET, Hi-C or eQTL)*". Thus, since JEME and GeneHancer used eQTLs and chromatin interactions to train their models, the result of the benchmarking procedure with those two datasets was likely biased and did not reflect the true situation. On the other hand, both eQTLs studies had the highest positive predictive values when EnhancerAtlas EGAs were used as a test dataset. Again, EnhanerAtlas enhancers were defined using the information about eQTLs, so it was not surprising that eQTLs were enriched among EnhancerAtlas data. Nonetheless, the highest percentage of overlaps with EnhancerAtlas could be also a result of the size of the benchmark dataset and the length of their anchors. For example, EnhancerAtlas anchors were up to 1Mb long and with this resolution it is hard to precisely identify interactions between regions that are usually in the order of a few hundred base pairs (Park 2009). In addition, both sources of chromatin interactions - the CCSI database and PC-HiC - had the highest positive predictive values when GeneHancer EGAs were used as a test datasets and as previously mentioned Fishilevich et al. (2017) used CHiC datasets from Mifsud et al. (2015) to predict EGAs. We expect that benchmarking results are biased for *flexibleC* and *stringentC* models as well, because they represent a consensus of five different sets of enhancer-gene associations (EGAs), including the GeneHancer and JEME. Therefore, *flexibleC* and *stringentC* models indirectly used the GTEx eQTLs and chromatin interactions as an input dataset to predict EGAs. This discrepancy was especially obvious when three *inhouse* datasets

were compared together: *inhouseM* (results of the reg2gene modelling of *gene expression* ~ *enhancer activity*) had consistently lower percentage of overlaps with tested benchmark datasets as compared to *stringentC* and *flexibleC* EGAs - sets of EGAs that represent a consensus between five different EGAs datasets including reg2gene *inhouseM* EGAs and which indirectly used eQTLs and chromatin interactions for their predictions. The rank of *stringentC* and *flexibleC* EGAs changed as different benchmark sets were utilized. Thus, it seems that only the result of the benchmarking procedure for FOCS EGAs was potentially unbiased in this analysis.

.....

Overall, we showed that, due to technical and biological limitations, benchmark datasets used in this analysis suffer from poor reproducibility. In addition, results of the benchmarking procedure are likely biased since eQTLs and chromatin interactions were initially used in most of the computational algorithms to predict enhancer-gene associations (EGAs). To account for this, we used 470 high-confidence *cis* enhancer-gene interactions assessed by cell-based CRISPR/Cas9 genetic screen from Gasperini et al. (2019) as a set of "positives" and considered predicted *cis* enhancer-gene interactions to represent true positives (TP), and unpredicted - false negatives (FN). In addition, we proposed an approach that searches for enhancer-gene "negatives" and assesses false positive (FP) and true negative (TN) EGAs. A proposed approach identified enhancer regions that were across all 127 cell types from the Roadmap Consortium (Roadmap Epigenomics Consortium et al. 2015) predicted by ChromHMM to be one of the repressed chromatin states (Ernst and Kellis 2012) and assumed that all corresponding enhancer-gene pairs should not be statistically associated. Overall, this enabled us, for the first time, to fully assess the performance of computational methods and assess all the elements of the confusion matrix: false positives (FP), false negatives (FN), true negatives (TN) and true positives (TP). Although *stringentC* had the smallest number of true positives (TP=45), it did not predict any FP. On the other hand, ~1,600 GeneHancer enhancer-gene associations were actually false positives, as well as ~36,000 EnhancerAtlas FP enhancer-gene associations. Strikingly, across all methods, *stringentC* and *flexibleC* had PPV over 0.9, but only *stringentC* reached the highest PPV of 1. Since *stringentC* and *flexibleC* models are actually ensemble classifiers of five different models of enhancer-gene associations (JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), FOCS (Hait et al. 2018), and reg2gene *inhouseM*), and we showed that they indeed predict more stringent EG associations, we corroborated the previous idea that the an ensemble classifier can be used to combine the classification rules of multiple classifiers to narrow down the

hypothesis space (Yang et al. 2010). We suspect that with this voting procedure, we "eliminated" all publication-specific enhancer-gene associations and retained only the robust ones. In addition, a difference in PPV between *stringentC* and *flexibleC* models indicated that more stringent voting procedure (more votes) predicted more precise results.

Nonetheless, because of the fact that reg2gene models used Roadmap datasets as an input, benchmarking using ChromHMM-predicted "negatives" from the Roadmap datasets could be, similar to benchmarking with eQTLs and chromatin interactions, biased. If that would be the case, JEME and FOCS, should have much higher PPV since Cao et al. (2017) and Hait et al. (2018) used Roadmap datasets as well to train their models. However, that was not the case. Thus, we suggest that, based on the currently available data, *stringentC* models are currently the most stringent set of computationally modelled enhancer-gene associations. In the future, when larger sets of experimentally validated true positive and true negative E-G pairs will be available, this can be a subject of further analysis.

5.5. Conclusions

A large set of experimentally validated true positive and true negative ETG pairs still lacks in the field of genomics. Nonetheless, *stringentC* is an interesting source dataset whose predictions are more precise and stringent than predictions of JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), FOCS (Hait et al. 2018) and it can be used to, for example, annotate non-coding SNPs to genes they regulate. This will be investigated in the last chapter of results of this thesis.

6

Results IV - Application of enhancer-gene associations in disease genetics: Stories of the GWAS Catalog, colorectal cancer and rs10411210

Preface

In this chapter, I used sets of predicted enhancer-gene associations to link non-coding polymorphisms to their putative causal genes. I was intrinsically motivated to perform this analysis.

Abstract

The assignment of non-coding risk SNPs to their target genes is not straightforward and it has been commonly approximated by linking SNPs to their proximal or eQTLs-associated genes. Here, we used enhancer-gene associations (EGAs) to annotate risk SNPs to their target genes by simply assigning SNPs to the genes associated with SNP-overlapping enhancers. We annotated three sets of non-coding SNPS: risk SNPs reported in the GWAS Catalog, colorectal cancer (CRC) associated SNPs and the rs10411210 with seven sets of enhancer-gene associations: JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), EnhancerAtlas 2.0 (Gao et al. 2016), FOCS (Hait et al. 2018), stringentC, flexibleC and inhouseM.

We identified that sets of annotated genes varied in their size - for example, using the EnhancerAtlas EGAs we associated ~400 genes to 312 colorectal cancer polymorphisms reported in the GWAS Catalog. On the other hand, the stringentC method predicted only 14 CRC genes, however, 65% of them were previously identified to be associated with the CRC. Likewise, results of downstream and functional analyses differed when SNPs were annotated to genes using different sets of EGAs. We analyzed a single SNP - rs10411210 - and identified that it could be annotated to 1-8 genes (and 18 EGAs). Likewise, some well-known CRC genes were identified by only certain sets of EGAs. Altogether, this suggested that we cannot easily assert one EGAs method to be a better source of predictions than others, since each of them can potentially add a piece of information needed to improve our understanding of genetic susceptibility. In addition, we showed that EGAs predict genes that were not previously reported for CRC, but were reported for some of the CRC-ancestral diseases. In addition, we show that EGAs can be potentially useful to identify novel SNP-gene associations, perform functional analysis of the underlying enhancer regions, and detect enhancer pleiotropy.

6.1. Introduction

Thousands of genomic polymorphisms have been statistically associated with human phenotypes and diseases through the genome-wide association studies (GWAS; Hindorff et al. 2009). The precise molecular mechanisms by which those polymorphisms exert their effects remains mostly unknown (Pickrell 2014), especially since the great majority of non-coding SNPs does not directly change the gene function or level of its products. Thus, non-coding risk SNPs are considered to modulate disease etiology by causing changes in the gene expression of a critical gene (Gerasimova et al. 2013).

The importance of enhancers in disease etiology, especially cancer genomics, was long under-estimated (Rheinbay et al. 2017; Chen et al. 2018). Putative causal genes were commonly annotated to SNPs based on their proximity to genes (Welter et al. 2014) and/or overlap with eQTLs (GTEx Consortium et al. 2017). As the role of enhancers in genetic susceptibility to various human traits and diseases became more evident (Smith and Shilatifard 2014; Chen et al. 2018), SNPs started to be more frequently annotated to their putative causal genes based on the overlap with regulatory regions (Styrkarsdottir et al. 2018; Short et al. 2018; Zhang et al. 2018; Schork et al. 2019). However, due to an ever-growing number of approaches developed to estimate enhancer-gene associations and/or sources of putative EGAs, researchers have been flooded by immense amounts of available information. For example, several methods that integrate epigenetic and genetic information have been developed: RegulomeDB (Boyle et al. 2012), HaploReg (Ward and Kellis 2012), FunciSNP (Coetzee et al. 2012), GWAS3D (Li et al. 2013), rSNPBase (Guo et al. 2014); and multiple databases such as EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017) can be used to annotate SNPs.

Nonetheless, guidelines on how to make choices while running the GWAS downstream analyses do not exist now. There are no recommendations that describe which source of data is relevant for a specific disease or SNPs context. In addition, a systematic review that compares all sources of information in the context of GWAS was never conducted. In this chapter, I set my focus on reviewing the performance of various EGAs methods to annotate risk polymorphisms from the GWAS Catalog, colorectal cancer (CRC) SNPs, rs104111210.

6.2. Methods

6.2.1. The intuition behind the hierarchical SNP annotation protocol

I annotated SNPs to putative genes hierarchically. Each SNP was first tested for the overlap with the promoter region. SNP was annotated to the nearest gene if an overlap with promoter region (+/-1000bp from the TSS) was identified. The remaining genes were tested for the overlap with enhancer regions, and annotated to the enhancer-associated gene. Lastly, genes remaining from the second step were assigned to the nearest gene within 1Mb distance. This procedure was repeated separately for each set of EGAs (**Figure 6.1.**).

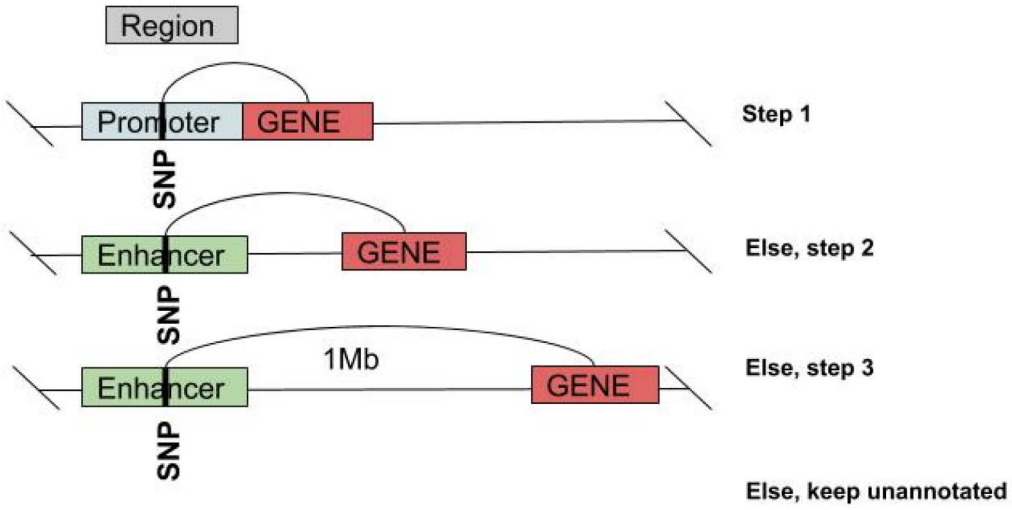


Figure 6.1. Schematic representation of the hierarchical SNP annotation protocol: in step 1, promoter-overlapping SNP is annotated to the nearest gene. Else, in step 2, enhancer-overlapping SNP is annotated to the enhancer-associated gene, Else, step 3: annotate SNP to the nearest gene that is located +/-1Mb, else keep the SNP unannotated.

6.2.2. An overview of used datasets

As a source of information of disease associated genetic polymorphisms, I selected two databases: the GWAS Catalog (Welter et al., 2014) and DisGeNET database (Piñero et al., 2017).

The GWAS Catalog

The GWAS Catalog (Welter et al., 2014) is a free online database that compiles unstructured data of genome-wide association studies and summarizes it into easily accessible high quality data.

DisGeNET

DisGeNET database is one of the largest publicly available collections of genes and variants associated with human diseases (Piñero et al., 2017). In July 2019, it contained 310,502 unique entries (after I excluded all associations reported in the GWAS Catalog or GWASDB) with 10,012 genes and 20,607 diseases.

Benchmarking datasets

As "CRC benchmark" gene sets, we selected three sets of CRC genes reported in the literature:

1) 63 CRC genes reported in the most comprehensive review of genetic susceptibility to colorectal cancer (Peters et al., 2015),

2) 1,676 genes reported in the DisGeNET database (Piñero et al., 2017) - one of the largest and comprehensive collections of human gene-disease associations - to be associated with colorectal cancer, and

3) 2,096 genes from the DisGeNET database that were associated either with colorectal cancer or any ancestral disease to CRC: intestinal cancer, carcinoma, cancer, neoplasm. I further refer to this gene set as *DisGeNET_A*.

6.3. Results

6.3.1. The GWAS Catalog in numbers

In February 2019, the GWAS Catalog contained 61,574 unique, non-coding SNP-disease associations (corresponding to 17,724 genes and 2,980 traits/diseases). The highest number of polymorphisms was reported for blood protein levels (N=2,334). Schizophrenia and breast cancer were associated with the highest number of non-coding polymorphisms (N=956 and N=755, **Figure 6.2.**).

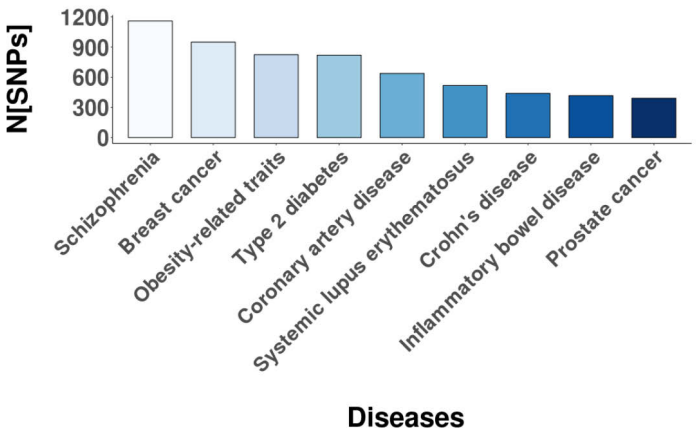


Figure 6.2. Histogram of a number of non-coding genetic polymorphisms (SNPs) reported per individual disease in the GWAS Catalog. GWAS reported traits (not diseases) were excluded from this figure.

We determined the nearest gene for each GWAS SNP and included only those genes that were located +/- 1Mb around the SNP position. We obtained a list of 18,003 genes (I did not manage to annotate only 47 SNPs). With this procedure we identified that 42.4% of the GWAS genes (reported or mapped) corresponds to the nearest gene and 38% of entries do not have reported gene associations - SNPs were associated with either “intergenic”/“genic” regions or reported to have unknown gene association (NR).

To analyze putative differences that could result when SNP-gene links would be assessed using different sets of EGAs, we performed SNP annotations using seven sets of EGAs: EnhancerAtlas

(Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018), *stringentC, flexibleC and inhouseM*. First, we excluded 3,661 GWAS Catalog SNPs that overlapped with promoter regions, and thus, were annotated to the promoter-regulating genes. This number of associations corresponds to 0.8% of the total SNP-gene associations assessed using EnhancerAtlas, or 4% and 4.1% of associations assessed using the *stringentC* models or FOCS (**Figure 6.3.**).

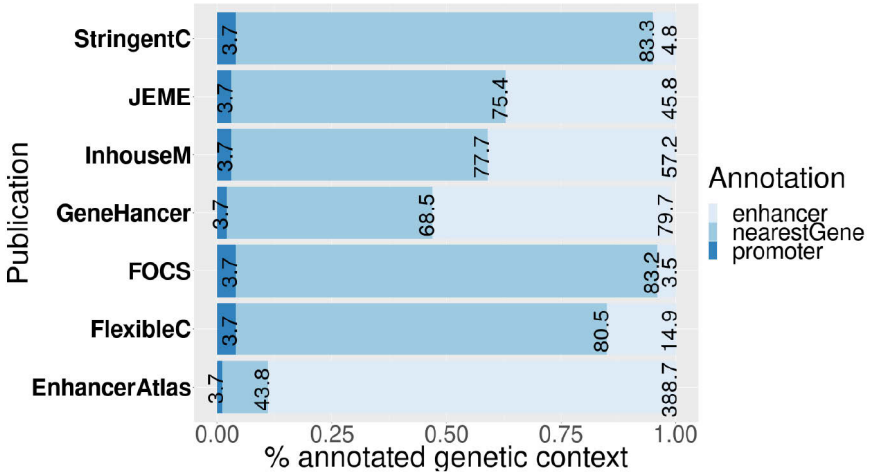


Figure 6.3. Percentage and number of the GWAS Catalog SNPs annotated to genes. Different colors separate SNPs that were annotated to genes based on the overlap with promoter, enhancers or minimal distance to the nearest gene. Percentages of overlap in each category can be seen on the x-axis. Reported numbers correspond to the total number of GWAS SNPs that were annotated in each of the three categories (reported numbers should be multiplied with one thousand to get the true number). For example, a total of 3,661 SNP was annotated based on the overlap with promoters (N=3.7), which corresponds to 0.8% of EnhancerAtlas-based SNP annotations. On the other hand, 388,702 EnhancerAtlas-based SNP annotations (or 89.1%) were done based on overlap with EnhancerAtlas enhancers. Seven sets of enhancer-gene associations used to annotate GWAS SNPs are indicated on the y-axis.

6.3.2. Results of the SNP-to-gene annotation analysis differ if distinct enhancer-gene associations (EGAs) are used to annotate SNPs - the GWAS Catalog

In the next step, we exclusively focused on non-coding SNP annotations. More than 89% (N=388,702) of SNP-gene associations could be assessed based on the overlap between SNPs and EnhancerAtlas enhancer regions. This corresponded to a four-time increase in the number of SNP-gene(s)-disease associations as compared to the original GWAS Catalog (**Figure 6.4., Supplementary Table 3.**). Likewise, the initial 61,574 GWAS Catalog SNP-gene(s)-disease associations increased to 79,680 if the GeneHancer EGAs were used as a source of information.

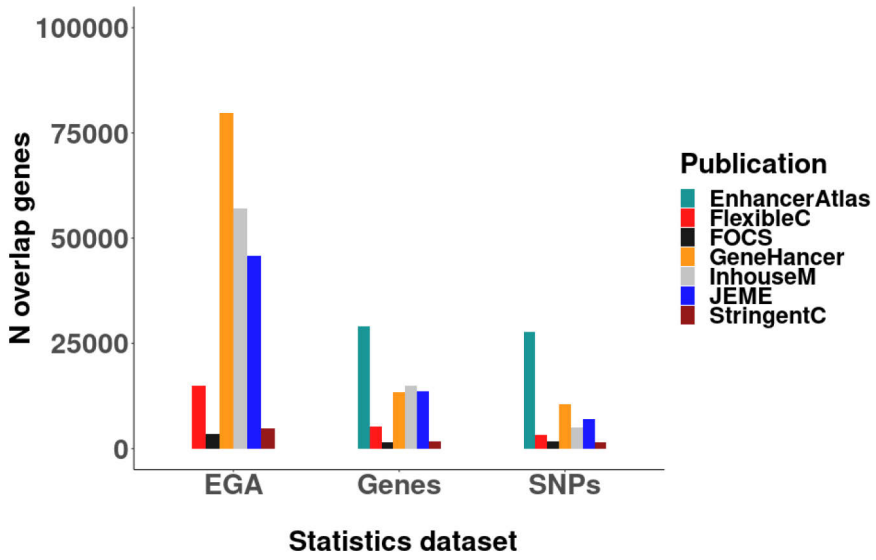


Figure 6.4. Histogram of the general statistics of the GWAS Catalog SNP annotation analysis based on seven enhancer-gene association (EGA) sets: JEME (Cao et al., 2017), GeneHancer (Fishilevich et al., 2017), EnhancerAtlas (Gao et al., 2016), FOCS (Hait et al., 2018), *stringentC*, *flexibleC* and *inhouseM.* Histograms indicate numbers of SNPs, genes and EGAs that were assessed in the annotation procedure. Largest EGA datasets consistently annotated largest numbers of SNPs, genes and EGAs. For example, using GeneHancer EGAs GWAS Catalog SNPs were associated with ~25K enhancers, and ~75K enhancer-gene associations

However, the number of gene-disease pairs decreased for other annotations: using the *flexibleC* we managed to annotate 14,857 SNPs (15%), whereas using the *stringentC* EGAs we linked 4,756 SNPs with genes via overlap with their enhancers (5.2%). Only 3,535 entries from the GWAS Catalog overlapped FOCS enhancers.

For each EGA method, we identified a reduction in the number of enhancers as compared to SNP, which implied that multiple SNPs overlapped a single enhancer. For example, 10,487 GeneHancer-annotated GWAS Catalog SNPs overlapped 8,601 enhancers. However, for EnhancerAtlas we observed the opposite trend, SNPs were commonly annotated to more than one enhancer: 27,769 EnhancerAtlas-annotated GWAS Catalog SNPs overlapped 70,718 enhancers (**Supplementary table 3.**).

6.3.3. Results of the SNP-to-gene annotation analysis differ if distinct enhancer-gene associations (EGAs) are used to annotate SNPs - colorectal cancer and rs10411210

As we confirmed our expectation that results of the SNP-to-gene annotation analysis differ if distinct enhancer-gene associations (EGAs) are used to annotate SNPs from the GWAS Catalog, we tested whether that holds true in smaller samples of risk-associated SNPs. First, we annotated

colorectal cancer (CRC) associated SNPs with seven sets of EGAs. Importantly, colorectal cancer has been a subject of an extensive research in our lab (Ronen et al., 2019). It has a large genetic component - heritable factors contribute to around 35% of the variation in risk of developing CRC (Jiao et al., 2014), genes and pathways that are important for initiation and progression of colorectal cancer were previously identified (Fearon, 2011; Ronen et al., 2019), and some of the associated genetic polymorphisms were confirmed experimentally (Goss and Groden, 2000). In addition, the majority of genetic variation that conveys the risk of CRC is located in non-coding genomic regulatory regions and their gene targets are unknown (Timofeeva et al., 2015).

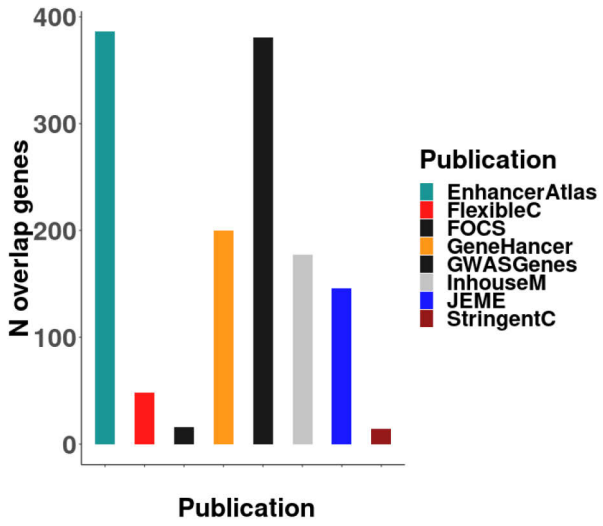


Figure 6.5. Histogram of the number of gene sets that were found to be associated with colorectal cancer SNP. Analysis was based on the annotation of 312 CRC SNPs reported in the GWAS Catalog using an overlap between SNPs and enhancers from seven enhancer-gene association sets (details of analysis explained above). Number of genes associated with the GWAS Catalog was reported as well.

We annotated 312 colorectal cancer polymorphisms reported in the GWAS Catalog (Welter et al., 2014). Although 381 CRC genes were reported in the GWAS Catalog, the largest EGA dataset - EnhancerAtlas - identified the largest number of entries: EnhancerAtlas predicted up to 386 CRC genes, GeneHancer around 200, whereas the *flexibleC* identified 48 CRC genes and *stringentC* method predicted only 14 CRC genes (**Figure 6.5., Supplementary Table 4**).

We additionally performed the pairwise comparison of CRC annotated gene sets and identified that the two largest datasets: EnhancerAtlas and GeneHancer had the largest gene overlap (**Figure 6.6.**). On the other hand, as the *stringentC* method represents a consensus of different EGA methods, gene set predicted with this method had the highest percentage overlap with other sets of predicted genes.

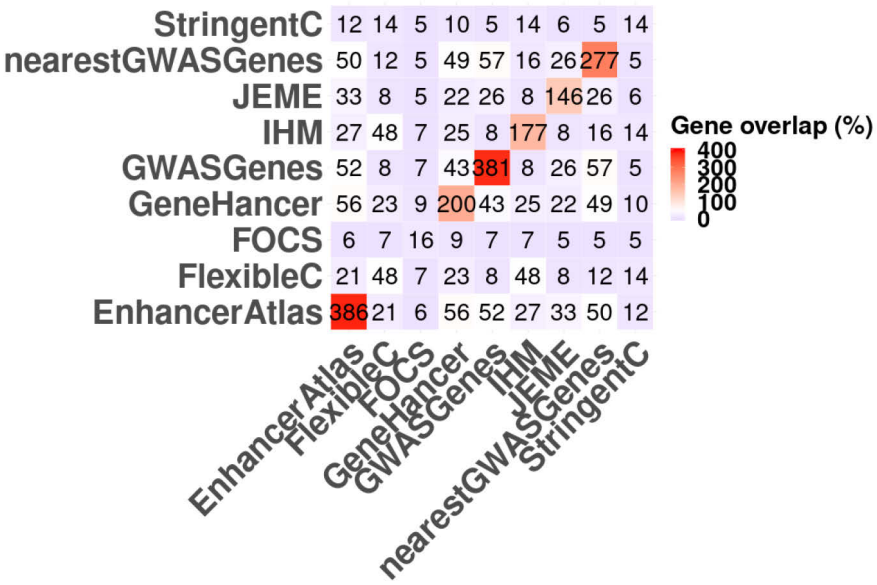


Figure 6.6. Heatmap of the pairwise gene overlap between CRC-associated sets of genes. Associations were identified based on the overlap between CRC-reported SNPs and enhancers from seven enhancer-gene association sets: EnhancerAtlas (N=386, Gao et al. 2016), JEME (N=146, Cao et al. 2017), GeneHancer (N=200, Fishilevich et al. 2017), FOCS (N=16, Hait et al. 2018), *stringentC, flexibleC and inhouseM.* In addition, CRC genes from the GWAS Catalog (N=381) and genes nearest to the SNP were reported (N=277).

A total of 16 genes was confirmed by more than 3 sets of EGAs, but all seven methods agreed upon CRC association for only two genes: RHPN2 and SERPINH1. Genes such as RPS3, KLF13 or GDPD5 were not associated with CRC only when FOCS EGAs were used as a source of information (**Figure 6.7.**).

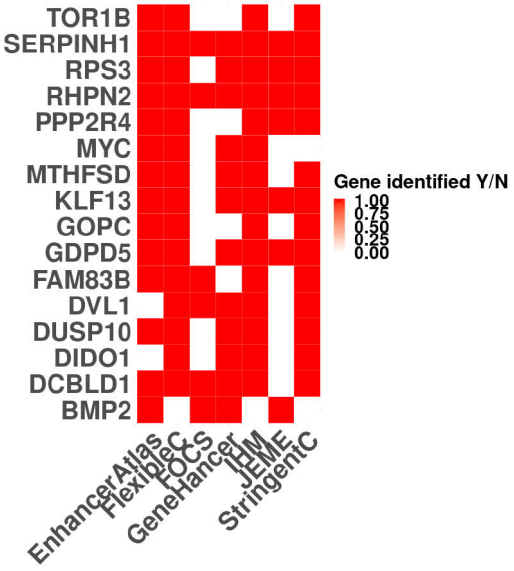


Figure 6.7. Presence of absence of a specific gene in a set of CRC-associated genes as assessed by seven different methods (x-axis). Red color marks the presence of a gene, whereas white marks its absence from the analyzed set. Only association of RHPN2 and SERPINH1 genes were corroborated across all sets of EGAs

Next, we selected one of the CRC-associated non-coding SNPs - rs10411210 - and annotated it with seven sets of EGAs (**Figure 6.8.**). Three fundamental pieces of information motivated us to study rs10411210. First, intronic enhancers were commonly found to be engaged in the long-range gene interactions with distant genes (Pomerantz et al. 2009; Harismendy et al. 2011; Maurano et al. 2012; Smemo et al. 2014). Second, along with cancer-associated genes, clusters of aberrantly active gene enhancers that drive dysregulated expression of oncogenes were previously identified in many cancer types (Sur and Taipale 2016). Third, its link to a region with regulatory potential was previously suggested and enhancer-correlated histone modifications were enriched at the 19q13.1 locus (Carvajal-Carmona et al. 2011). However, the previous bioinformatics analysis, which systematically searched for enhancer elements at this loci, was unable to pinpoint the cancer-causing element (Niittymäki et al. 2011).

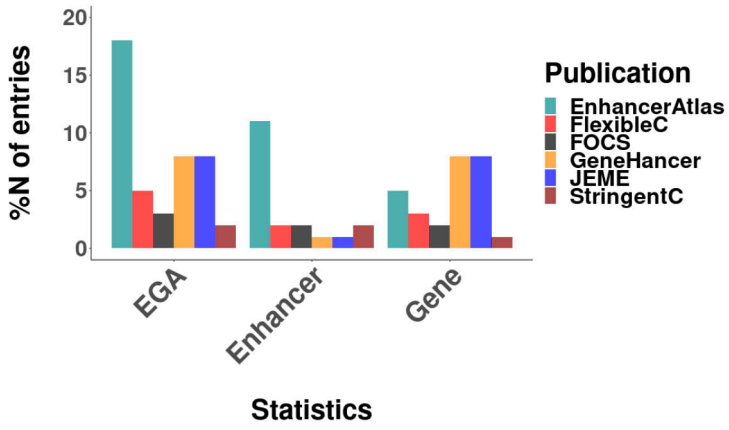


Figure 6.8. General statistics describing enhancer and gene annotations identified for the rs10411210 SNP. SNP annotations were obtained using seven sources of enhancer-gene association sets: EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018), *stringentC, flexibleC and inhouseM* by overlapping rs10411210 with enhancer regions. For example, eleven EnhancerAtlas enhancers, five genes and 18 EG pairs overlapped rs10411210.

In general, we identified eleven enhancer regions that were associated with rs10411210 when annotations were assessed using EnhancerAtlas EGAs. Those enhancers were paired with five genes and participated in 18 EnhancerAtlas enhance-gene associations (**Figure 6.9.**, **Supplementary Table 5**). Other methods (except for GeneHancer and JEME) annotated rs10411210 to more than one enhancer as well. *stringentC* annotated one gene to two enhancer

regions: one larger - chr19:33531704-33533702 (1999bp) and one shorter 439 bp (chr19:33532126-33532564) which we further researched in detail**.**

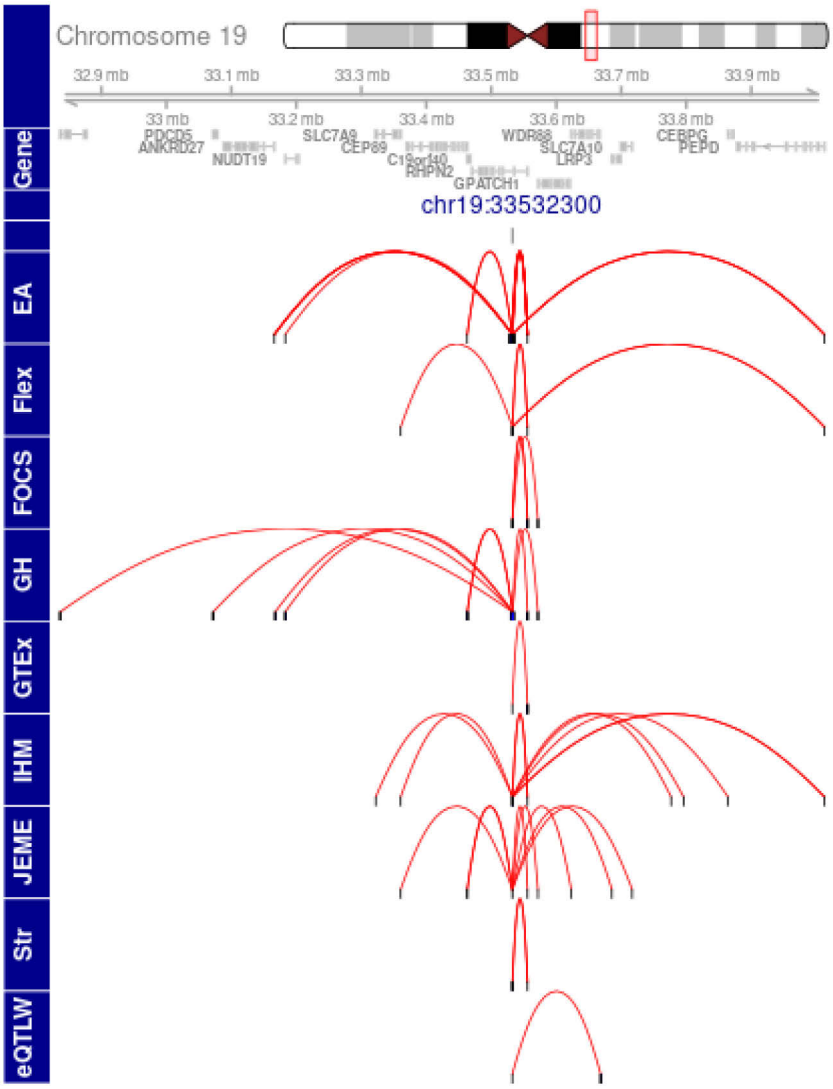


Figure 6.9. A visual representation of the results that were obtained by annotating a single SNP - rs10411210 - using enhancer-gene association from seven publications/methods and eQTLs from Westra et al. 2013 and the GTEx database (GTEx Consortium et al. 2017). From one to eight genes were annotated to this SNP by overlapping it with enhancer regions. Seven sources of enhancer-gene association sets were: EA - EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GH - GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018), Str - *stringentC, Flex - flexibleC and* IHM - *inhouseM.*

In the previous studies, rs10411210 was directly annotated to the RHPN2 gene due to its location in the intronic region of this gene (COGENT Study et al. 2008, Carvajal-Carmona et al. 2011, Niittymäki et al. 2011). All seven methods confirmed this association as well (**Figure 6.10., Supplementary Table 6**).

However, we linked rs10411210 with 17 (16) other genes: CEP89, ANKRD27, PEPD, NUDT19, SLC7A9, GPATCH1, FAAP24, ZNF507, PDCD5, CTD-2085J24.3, CTD-2540B15.9, CEBPG, CTD-2540B15.12, C19orf40, WDR88, LRP3, SLC7A10. Although FAAP24 and C19orf40 actually represent the same gene (rs10411210 was annotated to FAAP24 by GeneHancer, whereas rs10411210 - C19orf40 link was reported by JEME), this association was missed by the *flexibleC* EGAs. This indicated it as a limitation of our annotation method.

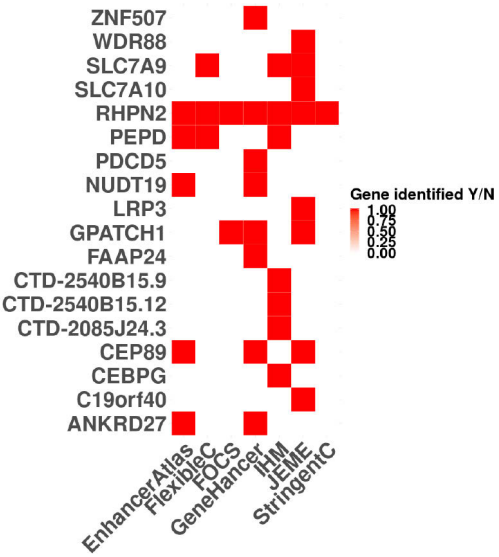


Figure 6.10. Presence of absence of a specific gene in a set of rs10411210-associated genes assessed by seven different methods (rows). Red color marks the presence of a gene, whereas white marks its absence from the analyzed set. Only association of the RHPN2 gene and rs10411210 is corroborated across gene sets

Other SNP-gene associations were rarely agreed upon. For example, SLC7A9 was confirmed only by JEME, *inhouseM* and *flexibleC* EGAs, whereas CEP89 was identified by EnhancerAtlas, GeneHancer and JEME. If GeneHancer EGAs would not be taken in consideration, we would miss association between CRC and PDCD5, whereas GPATCH1-CRC association was missed for *inhouseM* and EnhancerAtlas annotations.

All except for one association (chr19:33529636-33535923-ZNF507 from the GeneHancer database) were located within the same TAD region in the bowel cells (Dixon et al., 2012).

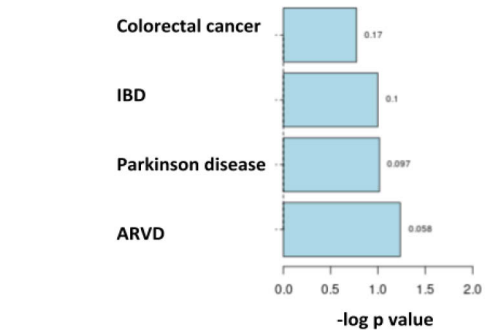
Since rs10411210 is only an index SNP - SNP assigned with the highest association significance with colorectal cancer in the underlying association study - we decided to identify which additional genes would be assigned to CRC if all SNPs in a haplotype block ~chr19:38 168 000–38 364 000 that contains rs10411210 itself would be analyzed (Carvajal-Carmona et al., 2011). We identified 21 SNPs that are in R2=0.8 LD with rs10411210 and successfully annotated 16 with all seven

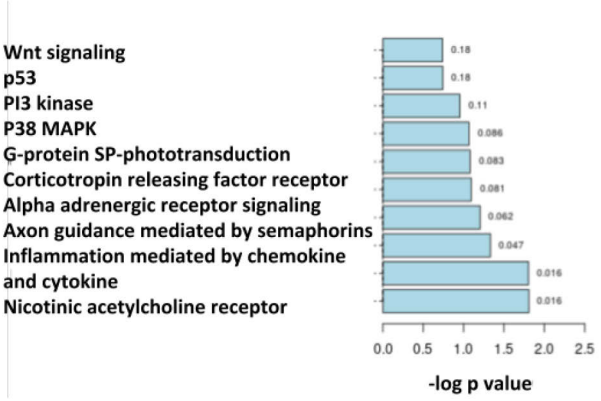
sources of EGAs (**Supplementary Table 7.**). As a result, we expanded the list of potential gene targets with six additional genes: KCTD15, CTD-2540B15.7, AC007773.2, CTD-2085J24.4, RN7SKP22, DPY19L3.

6.3.4. Results of the enrichment analysis differ if distinct enhancer-gene associations (EGAs) are used to annotate SNPs

Since we showed that results of the SNP-to-gene annotation analysis differ if distinct enhancer-gene associations (EGAs) are used to annotate risk SNPs, we further analyzed characteristics of identified sets of genes.

We performed the enrichment analysis of seven sets of CRC genes and obtained information for: molecular function, cellular component, biological process, human phenotype ontology, OMIM reported diseases (Kanehisa, 2002) and gene enrichment in the KEGG (Croft et al., 2011), Reactome (Hamosh et al., 2000), and PANTHER (Mi et al., 2013) pathways.





A. **B.**

Figure 6.11: Results of enrichment analysis for genes annotated to CRC SNPs. A. Results of the enrichment analysis (OMIM database) for 200 CRC genes annotated using GeneHancer EGAs. B. Results of the enrichment analysis (PANTHER pathways) for 146 CRC genes annotated using JEME EGAS. ARVD=Arrhythmogenic Right Ventricular Dysplasia (ARVD), IBD=inflammatory bowel disease.

As expected, gene sets showed enrichment for different pathways, phenotypes and diseases. For example, GeneHancer-annotated CRC genes showed the enrichment for colorectal cancer genes reported in the OMIM database (**Figure 6.11.A**) and well-defined CRC signaling pathways, TGFβ and Ras in PANTHER pathways (**Supplementary Figure 5.**). HPO (human phenotype ontology) did not indicate any enrichment for CRC, but did for sacral dimple, or abnormality of liposaccharide metabolism. Interestingly, JEME-annotated CRC genes were enriched for all previously identified CRC pathways (Wnt, PI3K-mTOR, Ras-ERK (MAPK), p53) in the PANTHER pathways database (**Figure 6.11.B**). On the other hand, the same set of genes was enriched in prostate cancer genes in OMIM and prostate neoplasm genes in HPO (**Supplementary Figure 5.**). Analysis of EnhancerAtlas-annotated CRC genes identified enrichments for cardiomyopathies (HPO, OMIM), phospholipid traits (GO_BP, biological processes) and RNA PolII activity (GO_MF, molecular function), whereas, *stringentC* gene sets showed enrichment for inositol-1,3,4,5-tetrakisphosphate-5-phosphatase activity, PI3K regulatory subunit binding, etc. (**Supplementary Figure 5.**).

6.3.5. Up to one fifth of the newly annotated CRC genes could be easily benchmarked

Next, we set off to benchmark results of the SNP-to-gene annotation analysis specifically for colorectal cancer associated SNPs. To achieve that we used three "benchmark" datasets: CRC genes reported in Peters et al., (2015) and DisGeNET database (Piñero et al., 2017), and genes in the DisGeNET database that are associated with CRC, intestinal cancer, carcinoma, cancer and neoplasm (*DisGeNET_A*).

Importantly, Peters et al. (2015) conducted the most comprehensive review of genetic susceptibility to colorectal cancer genes in which all genes previously associated with CRC were reported. The intuition behind defining the third benchmark dataset - *DisGeNET_A* - was to include

genes that have not yet been associated with CRC, but could likely be in the future. We hypothesized that some of the unknown CRC genes were associated with non-CRC diseases through GWASes and the majority of them, if truly associated with CRC, should be associated with CRC related-phenotypes such as the intestinal cancer, carcinoma, cancer, neoplasm.

We identified that EnhancerAtlas annotated CRC genes had the highest overlap with the CRC genes reported in the DisGeNET database (N=37/386). We could additionally link 11 genes with the CRC related-phenotypes in the DisGeNET database - previously, those 11 genes were not directly associated with CRC. (**Figure 6.12.**, **Supplementary Tables 8. and 9.**).

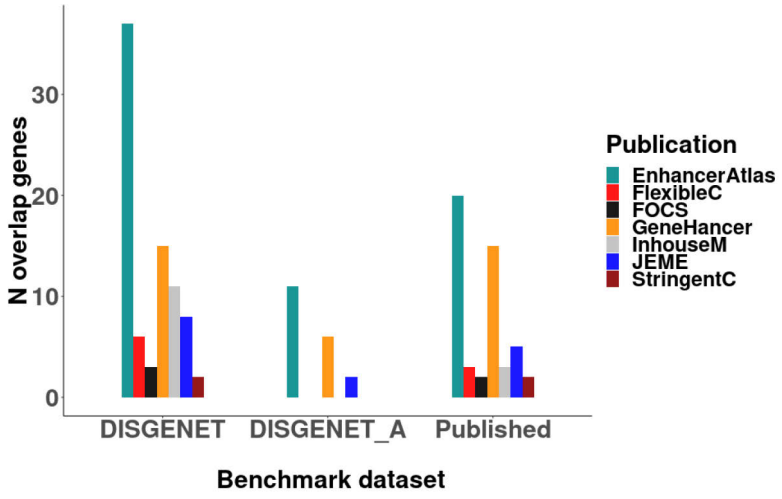


Figure 6.12. Histogram of the number of overlaps between three CRC benchmark gene sets and seven CRC gene sets that were identified by SNP-to-gene annotation procedure. Benchmark CRC gene sets correspond to: 1) DisGeNET - 1,676 CRC genes reported in the DisGeNET database, 2) DisGeNET_A - 2,096 genes from the DisGeNET database associated with colorectal cancer, intestinal cancer, carcinoma, cancer, neoplasm and 3) Published - 63 CRC genes reported in Peters et al. 2015. Seven CRC gene sets were assessed based on SNP-to-gene annotation procedure using EGA reported in EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018), *stringentC, flexibleC and inhouseM*. Importantly, only novel overlaps between sets of genes were reported for DisGeNET_A. For example, GeneHancer-annotated CRC genes overlapped 21 DisGeNET genes (15 DisGeNET and 6 *DisGeNET_A* CRC genes).

GeneHancer-annotated CRC genes overlapped 21 DisGeNET genes (15 CRC DisGeNET and 6 *DisGeNET_A* CRC genes). A total of 10 JEME genes overlapped DisGeNET genes (8 CRC and 2 additional *DisGeNET_A* CRC genes), whereas 3 FOCS genes were identified in DisGeNET, but we did not identify any novel gene-CRC association using the *DisGeNET_A* set of genes (**Figure 6.12.**). From the 14 CRC *stringentC*-annotated genes two (DUSP10 and DVL1) overlapped CRC genes reported in DisGeNET and two genes (DUSP10 and RHPN2) overlapped previously reported CRC genes (Peters et al., 2015). However, none of the identified genes associated with CRC was

additionally identified in the *DisGeNET_A* set of CRC genes (the same holds true for *inhouseM* and *flexibleC* gene sets).

If percentages were taken into account, a maximum of 19% genes form a given dataset (FOCS EGAs) could be confirmed by CRC genes in the DisGeNET database. In the case of the *stringentC*-annotated genes that corresponded to 14%, whereas 10% of the EnhancerAtlas CRC genes was benchmarked.

6.3.6. 65% of CRC genes annotated using the *stringentC* EGAs was previously associated with the CRC

We further focused on the smallest group of genes that were annotated to CRC SNPS - 14 genes identified using *stringentC* EGAs: DUSP10, MTHFSD, RHPN2, RS3, SERPINH1, GDPD5, KLF13, DCBLD1, GOPC, DIDO1, DVL1, FAM83B, PPP2R4, TOR1B.

Of those, only the RHPN2 gene is a well-described gene associated with CRC and reported in the GWAS Catalog (He et al., 2015; Tenesa and Dunlop, 2009). Nonetheless, we identified that seven other reported genes were associated with CRC as well: DUSP10 (Png et al., 2016), RPS3 (Tang et al., 2019), DVL1 (Kho et al., 2009), GOPC (Terasaki et al., 2002), DCBLD1 (He et al., 2016; Kang et al., 2015), GDPD5 (Feng et al., 2018) and KLF. Two genes (DUSP10 and DVL1) were the CRC genes reported in the DisGeNET database and two genes (DUSP10 and RHPN2) overlapped CRC genes reported in (Peters et al., 2015). Majority of the remaining *stringentC*-annotated CRC genes were previously implicated in cancer: PPP2R4 in posterior fossa group B ependymoma (Xing et al., 2006), FM83B in breast, lung, ovary, cervical, testis, thyroid, bladder, and lymphoid cancers (Cipriano et al., 2014), SERPINH1 in neuroblastoma, breast cancer progression, gastric and stomach cancer (Zhang et al., 2010); (Zhu et al., 2015), KLF13 in glioma (Wu et al., 2019), and DIDO1 in adenocarcinoma (Liu et al., 2010). MTHFSD was reported in the GWAS Catalog for FEV/FEC ratio and TOR1B for dystonia (Welter et al., 2014).

We studied whether we could discern that some of the above SNP-gene associations could be more biologically relevant by analyzing their co-localization in the same TAD region. For this analysis, we were inspired by work of Symmons et al., (2014) - they identified that enhancers are

generally contained within the same TAD as genes they regulate, thus interacting more likely (with higher frequency) than if they are contained in different domains. We hypothesized that the "co-localization" of SNPs (overlapping enhancers) and genes in the same TAD region could be used as the filtering feature to select the true SNP-gene associations: those that do not overlap the same TAD region are false positives, and those that are identified in the same TAD are true positives. We analyzed TAD regions reported in (Dixon et al., 2012; Nora et al., 2012, cells of the bowel tissue) and screened for the [SNP/enhancer]-gene pairs that were located within the same TAD region. Thirteen genes indeed were located in the same TAD region as their enhancer. However, DIDO1, PPP2R4 and TOR1B were not identified in the same TAD region as their SNP-overlapping enhancers (**Table 6.1.**).

Table 6.1. Genes found to be annotated to twelve CRC-associated SNPs from the GWAS Catalog. Their association with specific phenotypes was indicated, as well as co-localization in the same TADs in the bowel tissue (Dixon et al., 2012).

Gene	Associated phenotype	Within TAD
DUSP10	**Colorectal tumorigenesis** (Png et al., 2016)	yes
MTHFSD	FEV/FEC ratio (MacNair et al., 2016; Shrine et al., 2019), plasma parathyroid hormone levels (Matana et al., 2018), amyotrophic lateral sclerosis (MacNair et al., 2016; Shrine et al., 2019)	yes
RHPN2	**CRC-association** report for rs10411210 (He et al., 2015; Tenesa and Dunlop, 2009)	yes
RPS3	Changes in gene expression in **colon adenocarcinomas** and **adenomatous polyps** compared to adjacent normal colonic mucosa (Tang et al., 2019) and breast cancer (Ono et al., 2017)	yes
DIDO1	Esophageal adenocarcinoma (Ono et al., 2017)	**No**
PPP2R4	Posterior fossa group B ependymoma (Xing et al., 2006)	**No**
TOR1B	Dystonia, autoimmune	**No**
FAM83B	breast, lung, ovary, cervical, testis, thyroid, bladder, and lymphoid cancers (Cipriano et al.,2012; Cipriano et al., 2014; Okabe et al., 2015)	yes
DVL1	Wnt signal pathway in **colorectal cancer** (Kho et al., 2009)	yes
GOPC	prognostic marker in **colorectal cancer** (Terasaki et al., 2002), lung adenocarcinoma, angiosarcoma, etc. in the DisGeNET database (Piñero et al., 2017).	yes
DCBLD1	**colorectal cancer (He et al., 2016; Kang et al., 2015)**, squamous cell	yes

	carcinoma, bile duct cancer, lung cancer, adenocarcinoma, uterine corpus endometrial carcinoma, etc.	
SERPINH1	neuroblastoma, breast cancer progression, gastric and **stomach cancer** (Zhang et al., 2010); (Zhu et al., 2015)	yes
KLF13	glioma (Wu et al., 2019), cholesterol biosynthesis and **colorectal cancer** development (Yao, 2019), prostate cancer (Wang et al., 2018)	yes
GDPD5	CRC (Feng et al., 2018)	yes

6.3.7. Enhancer-binding TFs and co-factors were previously reported in colorectal cancer

We performed a more detailed analysis of rs10411210-to-gene annotation results to pinpoint the underlying biological mechanism of rs10411210 association with colorectal cancer. To achieve that we analyzed enhancer locations, transcription factor binding site (TFBS) motifs and CHiP-Seq signals for transcription factors. First, we compared locations and size of enhancers across methods (**Figure 6.13.**).

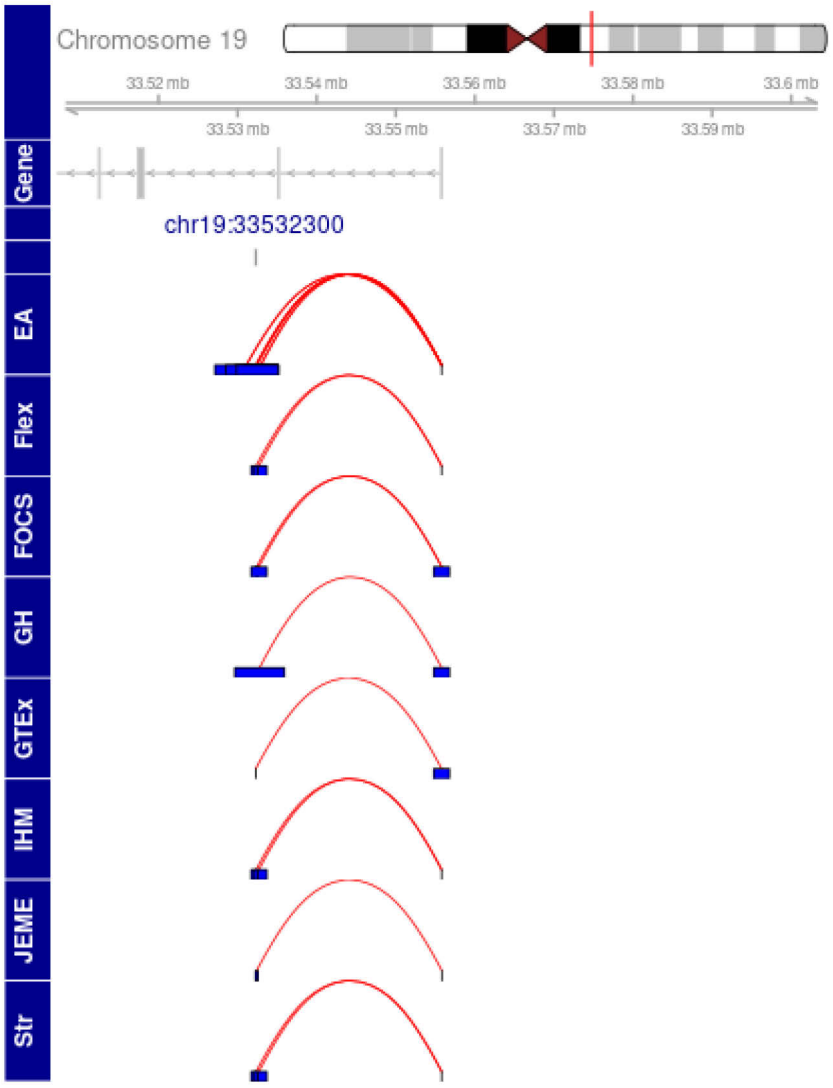


Figure 6.13. A visual representation of results that were obtained by annotating a single SNP - rs10411210 - using nice different approaches: enhancer-gene association from seven publications/methods and eQTLs from Westra et al. 2013 and the GTEx database (GTEx Consortium et al. 2017). From one to eight genes were annotated to this SNP by overlapping it with enhancer regions. Seven sources of enhancer-gene association sets were: EA - EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GH - GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018), Str - *stringentC, Flex - flexibleC and* IHM - *inhouseM.*

The size of enhancers varied from several hundred to several thousand base pairs: a maximum length of 8,031 bp was reported for EnhancerAtlas enhancers, whereas JEME and in-house models

report 345 and 439 bp enhancers. Thus, we could not precisely identify an enhancer location that conveyed the SNP-disease association. However, based on our previous knowledge about defined enhancer regions (*stringentC* enhancers are the most robust definition of enhancer regions), we decided to select and analyzed one enhancer region - chr19:33532126-33532564 - which corresponded to 439bp "consensus" enhancer region reported in *stringentC* models. For a given enhancer, we identified binding sites for cancer-associated TFs and TF-coactivators associated with colorectal cancer such as ATF3, CREB1, EP300 and STAT3 (**Supplementary Table 10.**). In addition, we identified TF binding motifs for CRC-related transcription factors: GATA2, SP1, STAT3, JUNB, SOX9, etc.

Initial bioinformatics analysis failed to pinpoint the cancer-causing element at the 19q13.1 locus (Niittymäki et al. 2011), but subsequent efforts found HEY1, RXRA, FOSL2, JunD, P300 and BAF155 to bind strongly to the region encompassing rs10411210 (Carvajal-Carmona et al. 2011). However, the identified CHiP-Seq peaks usually spanned more than 1kb in length as compared to the 439-bp regions identified in this research. In addition, it is well known that many cancer malignancies require oncogenic and/or tumor suppressor TFs for their survival, proliferation, and disease progression (Murakawa et al. 2016) and many oncogenic TFs and co-factors were identified to be over-activated in colorectal cancer (Peters et al. 2015; Nagaraju and Bramhachari 2017).

6.3.8. Path to follow: enhancer pleiotropy detection with SNP-enhancer-gene annotations

By focusing our analysis to a single set of SNP-gene annotations (originated from *stringentC* models), we made few observations that could represent interesting avenues of future research. We selected *stringentC*-based annotations due to the fact that *stringentC* associations are based on the *consensus* enhancer definition, characterized by the highest positive predictive value and had the lowest number of predictions (Chapter 3).

First, we observed that SNPs from various association studies can commonly overlap the same enhancer regions - LOXL1 was one of such genes (**Figure 6.14.**). LOXL1 was previously associated with multiple SNPs and phenotypes in the GWAS Catalog and with multiple enhancers (Welter et

al., 2014). Its association was reported for coronary artery disease, aortic root size, waist-hip ratio, joint mobility, hand grip strength, height, etc. (**Supplementary Figure 6.**).

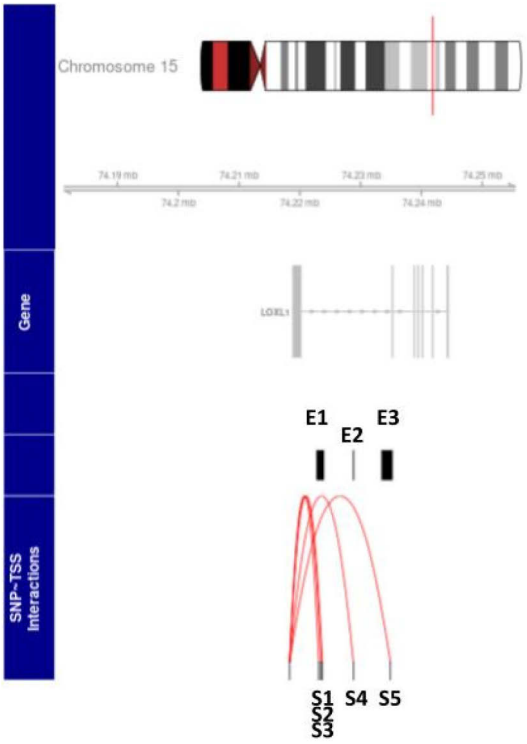


Figure 6.14. Visual representation of *stringentC* enhancers and GWAS Catalog SNPs that overlap with the LOXL1 gene. Three *stringentC* enhancer regions were found to be located in the intergenic regions of LOXL1 gene: E1 - chr15:74222688-74224043, E2 - chr15:74228665-7422857 and E3 - chr15:74233359-74235313. E1 overlaps with three GWAS Catalog SNPs associated with three different phenotypes: rs62004866 - hand grip strength, rs150025731 - joint mobility, rs28522673 - coronary artery disease. E2 regions overlap rs4886782 SNP associated with various waist circumference or BMI phenotypes. E3 enhancer and rs12441130 overlapping SNP is associated with heel bone mineral density

Only five out of twelve LOXL1-associated SNPs reported in the GWAS Catalog overlapped the *stringentC* enhancers, and thus, could be annotated. Three of them (rs62004866, rs150025731, rs28522673) overlapped one enhancer region: chr15:74222688-74224043 (E1); whereas rs4886782 overlapped chr15:74228665-7422857 enhancer (E2) and rs12441130 chr15:74233359-74235313 enhancer (E3). E1-overlapping SNPs were identified through three different association studies (GWAS): rs62004866 was associated with hand grip strength (Tikkanen et al., 2018), rs150025731 with joint mobility (Pickrell et al., 2016), and rs28522673 with coronary artery disease (van der Harst and Verweij, 2018)**, Supplementary Table 11.**). Thus, this example confirmed that SNPs associated with different phenotypes can overlap a single enhancer region.

Second, we identified that annotating SNPs-to-genes via EGAs has the potential to reveal novel SNP-gene-(disease) associations. We hypothesized that if we identified one SNP-gene association that was missed by GWA studies (but later confirmed), we could identify many others as well. For example, using the *stringentC* enhancer-gene associations (EGAs) we identified gene targets for a total of 4,756 GWAS Catalog SNPs, but 1,285 of those entries had no gene association reported (in the GWAS Catalog coded as: intergenic, Intergenic or NR - not reported). In addition, a total of 1,558 or 33% of interactions was confirmed by previously reported SNP-gene associations.

Specifically, we identified LOXL1 association with rs12441130. This polymorphism was previously associated with the heel bone mineral density phenotype and its association with the LOXL1 gene was reported in the GWAS Catalog (Kim, 2018). Nonetheless, we additionally identified the rs12441130 association with the LOXL1-AS1 gene (**Supplementary Figure 7.**), which was not identified through the GWA studies, but was recently confirmed by Pasutto et al., (2017). This example demonstrates that many novel SNP-gene associations could be identified by analyzing SNP-enhancer-gene links.

6.4. Discussion

Thousands of risk-associated, mostly non-coding, polymorphisms have been identified in the human genome (Welter et al. 2014). Only a handful of them have been functionally characterized or mechanistically linked to phenotypes (Frazer et al. 2009). Non-coding variants were shown to cause common diseases more likely than non-synonymous coding variants (Manolio et al. 2008) and they account for the vast majority of heritability (Gusev et al. 2014), however, their precise gene targets and the molecular mechanisms by which they exert their effects are mostly unknown (Pickrell 2014; Welter et al. 2014). Until today, risk SNPs have been generally annotated to their putative causal genes based on their proximity to gene targets (Welter et al. 2014) or by using information from the eQTL studies (GTEx Consortium et al. 2017). Consequently, more than 40% of genes reported (or mapped) in the GWAS Catalog correspond to genes that are closest to their risk SNPs (our in-house analysis).

Recently, the role of enhancers in genetic susceptibility to various human traits and diseases became more evident (Smith and Shilatifard 2014; Chen et al. 2018) and researchers started to annotate non-coding SNPs using information about enhancer-gene associations (EGAs) (Styrkarsdottir et al. 2018; Short et al. 2018; Zhang et al. 2018; Schork et al. 2019). Multiple sources of EGAs have been reported, however, we do not know if and how different sources of information influence results of annotation. To get a better understanding of this topic we annotated three sets of risk SNPs using seven sets of enhancer-gene associations: EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018) and our in-house models: *stringentC, flexibleC and inhouseM.* We simply overlapped GWAS SNPs; SNPs associated with colorectal cancer and rs10411210 with sets of enhancers and linked them to corresponding enhancer-associated gene(s).

We observed large differences in sets of genes annotated to all three sets of risk SNPs. Ten times more SNP-gene annotations would be observed if we used EnhancerAtlas EGAs as an annotation tool as compared to FOCS EGAs. This was likely a consequence of multi-gene SNP annotations - enhancers often regulate multiple genes (Maurano et al. 2012), and thus, more than one gene could be (and was) annotated to the majority of analyzed SNPs and enhancers. In addition, multiple EnhancerAtlas enhancers frequently overlapped a single SNP - especially obvious in the case of rs10411210 and its eight overlapping enhancers. In case of other methods, the majority of risk SNPs was actually located outside of defined enhancer regions.

We identified more than a ten-fold difference in the number of annotated genes across different EGA methods for 312 CRC SNPs and corroborated our previous findings for the GWAS Catalog. The largest set of enhancer-gene associations (EnhacerAtlas) identified the largest number of associated genes (~400). However, pairwise analysis of CRC-associated sets of genes did not show that any two sets of genes had a discernibly better overlap, although we did identify that all seven EGA annotations linked CRC SNPs with RHPN2 and SERPINH1. RHPN2 is a well-known gene associated with CRC (Tenesa and Dunlop 2009; He et al. 2015), whereas expression of SERPINH1, also known as heat shock protein 47 (HSP47), in colorectal cancer tissue was found to be significantly higher than in adjacent normal colonic mucosa (Mori et al. 2017).

Likewise, we identified differences between gene sets annotated to a single SNP - rs10411210. The link between rs10411210 and enhancer region was previously suggested (Carvajal-Carmona et al. 2011), but rs10411210 was consistently annotated to the RHPN2 gene (COGENT Study et al.

2008, Carvajal-Carmona et al. 2011, Niittymäki et al. 2011). We delineated 16 linked genes linked to this SNP, but only RHPN2 was confirmed by all methods. Other genes, such as GPATCH1 and CEP89 (identified by JEME), were previously reported for colorectal cancer; whereas PDCD5 and CEBPG genes were linked with either colorectal cancer or some other type of carcinoma in DISGENET or GWAS Catalog. Cohen et al. (2017) assessed GPATCH1 gene as a putative gene target for a recurrently gained variant enhancer loci that overlaps with rs10411210. Murakawa et al. (2016) reported that enhancers covering rs10411210 target the CEP89 gene. Likewise, ANKRD27, SLC7A10, SLC7A9, WDR88, ZNF507 were identified to be associated with cancer in the COSMIC database - the largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer (Forbes et al. 2011). Lastly, we identified an association between CEBPG, FAAP24 (C19orf40), LRP3, PDCD5 genes and cancer using an on-line text-mining tool - DISEASES (Pletscher-Frankild et al. 2015).

Importantly, some known CRC genes were identified by only a certain set of EGAs. This suggested that we could not easily assert one EGAs method to be a better source of predictions than others could. Especially since each of them can potentially add a piece of information needed to improve our understanding of genetic susceptibility. We analyzed whether we can single out some of the identified EGAs by analyzing their co-localization within the same TAD. However, we identified that only one EG association did not colocalize in TADs from the bowel cells (Dixon et al. 2012). Thus, TAD colocalization of enhancers and genes does not seem to be the most discriminative approach to differentiate between true and false enhancer-gene associations. However, it would be interesting to include additional TADs from other cell types and analyze whether certain EGAs are differentially found outside or inside of TAD regions in healthy versus individuals with colorectal cancer.

In addition, since rs10411210 is just an index GWAS SNP (COGENT Study et al. 2008, Zhang et al. 2014) at the 19q13.q locus, we studied the overlapping haplotype block at ~chr19:38 168 000–38 364 000 (Carvajal-Carmona et al. 2011) and identified 21 SNPs and six additional genes that are in LD (R2=0.8) with this SNP. Two of them: DPY19L3 and KCTD15 are reported in the COSMIC database (Forbes et al. 2011) to be associated with colorectal cancer (Lim et al. 2012).

.....

We used enrichment analysis to highlight differences across CRC-associated gene sets and identified that different gene sets showed enrichment for different pathways, phenotypes or diseases. Interestingly, only JEME-annotated CRC genes were enriched for all CRC associated pathways: Wnt, PI3K-mTOR, Ras-ERK (MAPK), p53. Since JEME predictions were assessed using the eQTL datasets (Cao et al. 2017) and eQTLs are very abundant across many databases including frequently used to discover novel pathways (GTEx Consortium 2015), the pathway analysis and JEME EGAs were assessed based on the same set of data which likely biased our analysis.

.....

By benchmarking sets of CRC-gene association with three sources of CRC-related genes: genes reported in Peters et al. 2015, DisGeNET database (Piñero et al. 2017), and genes associated with CRC, intestinal cancer, carcinoma, cancer and neoplasm in the DisGeNET database, we could confirmed up to one fifth of the newly annotated CRC genes (FOCS EGAs). Expectedly, the number of benchmarked genes correlated with the size of the EGA datasets. If we expanded the benchmark dataset with genes that are associated with diseases with similar etiology such as intestinal cancer, carcinoma, cancer and neoplasm with colorectal cancer, we managed to additionally benchmark some of the associated genes that were confirmed with targeted literature research. For example, although only RHPN2 is a well-described gene associated with CRC, the majority of other genes that were associated with colorectal cancer using *stringentC* EGAs could be confirmed as well (Png et al. 2016, Kang et al. 2015, He et al. 2016, Feng et al. 2018, Tang et al. 2019). Genes that have never been associated with CRC (DIDO1, PPP2R4 and TOR1B) "broke" the TADs boundaries - they were located outside of defined TADs. This indicated the potential of the TAD co-localization information to discern putative SNP-gene associations and corroborated the observations of Symmons et al. (2014) that the great majority of the genes colocalizes with their regulatory regions.

Using enhancer-gene associations reported in the *stringentC* models we identified some novel SNP-gene associations and opened novel avenues of our research. For example, we confirmed associations between rs12441130 and LOXL1 and LOXL1-AS1. The expression levels of LOXL1-E1A, LOXL1 and LOXL1-AS1 genes was shown to correlate with genotypes at the rs12441130 position and their alternative splicing is affected in hTCF cell line (Pasutto et al. 2017). However, LOXL1-AS1 - rs12441130 association represented a novelty (from the perspective of the GWAS Catalog). Thus, with this example, we demonstrated that many of the currently unknown annotations could

be discovered if the risk SNPs would be systematically analyzed in the light of the *stringentC* and other enhancer-gene associations.

Lastly, we analyzed the scope of information that can be revealed by performing a functional analysis of a single enhancer region chr19:33532126-33532564. This region overlaps CRC-associated SNP rs12441130. We identified a number of binding sites for cancer-associated TFs and TF-coactivators such as ATF3, CREB1, EP300 and STAT3 and the TF binding motifs for GATA2, SP1, STAT3, JUNB, SOX9, etc. Many of the aforementioned TF were associated with CRC: for example, SP1 is an important transcriptional regulator that plays a significant role in CRC initiation and metastasis (Bajpai and Nagaraju 2017). Reduced expression or overexpression of GATA transcription factors has been associated with CRC, whereas expression of the SOX9 gene was increased in CRC tissues compared with adjacent normal tissues (Lü et al. 2008). *In vitro* and *in vivo* studies showed that ATF3 promotes growth and metastasis of colon cancer tumors (Hackl et al. 2010) and AP-1 TF family members are differentially expressed in neoplastic and nonneoplastic colorectal tissues, whereas upregulation of Fra-1 and c-Jun represents an early event in human CRC tumorigenesis (Zhang et al. 2005). Nonetheless, additional experimental analyses should be performed to identify the underlying mechanism and TFs that mechanistically associate this with colorectal cancer.

We observed that SNPs from various association studies can overlap a single enhancer region. We found this information to be very exciting, since it could be used to identify novel examples of enhancer-based pleiotropy (Sabarís et al. 2019). In other words, changes in the same enhancer region can trigger different phenotypes and cause pleiotropy. For example, we identified locus chr15:74222688-74224043 (E1) a single intronic enhancer region within the LOXL1 gene that was associated with multiple SNPs and phenotypes in the GWAS Catalog: hand grip strength, joint mobility and coronary artery disease (Pickrell et al. 2016; Tikkanen et al. 2018; van der Harst and Verweij 2018). We speculate that if the link between aforementioned phenotypes really exists, it could be rheumatoid arthritis (RA) - a complex multisystem inflammatory disease characterized by the loss of immunologic self-tolerance, chronic inflammation, and destruction of the joints (McInnes and Schett 2011). Its association with an increased prevalence of coronary heart disease and a high cardiovascular (CV) mortality was previously identified (Goodson 2002; Gonzalez-Gay et al. 2005).

This approach can also be used to increase the number of testable variants associated with CRC as well. For example, the E1 enhancer region was found to be associated with two additional genes: ISLR and STOML1. For example, LOXL1 catalyzes the polymerization of tropoelastin to form the mature elastin polymer, allow efficient elastin core cross-linking and in knockout mice, it triggers a phenotype characterized by an abnormal elastic tissues and basement membranes of blood vessels (Liu et al. 2004). The STOML1 gene encodes stomatin - an integral membrane protein that localizes to the cell membrane of red blood cells and other cell types (Mairhofer et al. 2009). ISLR, as a member of the immunoglobulin superfamily, is likely involved in adhesion or binding to other proteins in solution or at the cell surface and implicated in immunity (Nagasawa et al. 1999). Nevertheless, a more detailed analysis is required to understand the link between the LOXL1 gene intronic enhancer and associated phenotypes. For the sake of writing this thesis, I did not dive deeper in the analysis of this link. However, it is a good showcase that demonstrates the potential of pooling information across different studies and discovering novel examples of enhancer-based pleiotropy. In addition, this could support Corradin et al. (2014) that different SNPs in the same LD block could identify enhancers that cooperatively regulate the same gene.

In this specific analysis, we were limited by the fact that we used only the *stringentC* enhancer-gene annotations. Thus, we could easily miss many other (potential or confirmed) gene associations. For example, a known association of the MYC gene and colorectal cancer (Peters et al. 2015) was confirmed by EnhancerAtlas, GeneHancer, *flexibleC*, or the *inhouseM* EGAs, whereas the PGC gene reported only in EnhancerAtlas is known to be involved in the PGC-1/ERR signaling axis in cancer (Deblois et al. 2013; LeBleu et al. 2014). Thus, researchers should start the SNP annotation analysis with the smallest set of enhancer-gene associations (*stringentC*) and then expand it with other available information. Likewise, although we included information from the HiC experiments (Dixon et al. 2012), we did not reflect on other possible sources of data that could provide us with additional information about SNP-gene-disease interactions such as eQTL studies, CRISPR-Cas experiments, etc. We did not include information about SNPs that are in LD with CRC-associated index SNPs as well.

6.5. Conclusions

We showed that many non-coding SNPs can be annotated to their target genes by “borrowing” information from enhancer-gene associations (EGAs) - we simply assigned SNPs to the genes associated with SNP-overlapping enhancers. Nonetheless, annotated sets of genes do differ when

different sets of EGAs are used as a source of information. Many established gene-disease associations are not necessarily present across all datasets and one needs to integrate information across datasets to be able to get a comprehensive understanding of gene regulation in health and disease.

7

Discussion

7.1. Summary

Complex organisms developed multiple mechanisms to regulate gene expression, but most regulation is believed to occur at the level of transcription initiation by *cis*-regulatory sequences that recruit a distinct set of *trans* factors. Proximal promoters (Lenhard et al. 2012) and distal enhancers (Lettice et al. 2003) are among the best-characterized *cis*-regulatory sequences in the human genome (Andersson et al. 2015). Promoters are located nearby the transcription start sites of genes and integrate a total regulatory input into the rate of transcriptional initiation (Lenhard et al. 2012), whereas enhancers are distal elements that interact with promoters and further refine gene expression across cell types and developmental stages (Banerji et al. 1981, Gerster et al. 1986, Blackwood and Kadonaga 1998).

How do enhancers spatially and temporarily modulate gene expression represents one of the central questions of genomics (Weber and Schaffner 1985; Gerster et al. 1986; Szutorisz et al. 2005,Muse et al. 2007; Zeitlinger et al. 2007; Core et al. 2008). Previously, low-throughput experiments empowered us to learn certain aspects of enhancer-mediated gene expression regulation; however, to fully understand and appreciate its complexity it is necessary to systematically identify and characterize all regulatory elements in a genome-wide manner and discern their cell-type specific patterns. Up today, multiple approaches have been utilized to study enhancer-mediated long-range gene regulation and they can be broadly categorized into four categories: predictions using information from the eQTL studies (Rockman and Kruglyak 2006; Gaffney et al. 2012; GTEx Consortium et al. 2017) or (3C)-derived technologies (Dekker et al. 2002; Simonis et al. 2006; Dostie et al. 2006; Lieberman-Aiden et al. 2009; Fullwood et al. 2009), and direct functional confirmation of enhancer activity by reporter assays or cellular screens (Arnold et al. 2013; Kwasnieski et al. 2012; Arnold et al. 2013; Kheradpour et al. 2013; Kvon 2015; Gasperini et al. 2019). However, they are hindered by technological and biological limitations of high-throughput technologies that are reflected in a high number of false positives and negatives (Hariprakash and Ferrari 2019).

The fourth approach, computational modelling of *gene expression ~ enhancer activity*, has been, as well, commonly used to map enhancers to their putative genes; and it was the main subject of this thesis. Recently, several data integration approaches aimed to predict enhancer-gene associations were developed using a large number of tissues, cell types and cell lines:

EnhancerAtlas (Gao et al., 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al., 2017), FOCS (Hait et al. 2018), HANCER (Wang et al. 2019). I reviewed and characterized differences between their results and I set off to develop a novel computational approach. I performed a thorough benchmarking of seven sets of computationally identified enhancer-gene associations and reviewed how different predictions of EGAs can influence the result of functional analysis of risk SNPs.

In **Chapter 3**, I systematically compared computational predictions of enhancer-gene associations and identified their differences. I showed that individual sets of predictions differ tremendously; especially in the location, number and properties of defined enhancer regions. Thereby, we showed (and confirmed) that the key bottleneck in understanding enhancer-promoter communication has been the lack of the tools to precisely identify and characterize the functions of large numbers of enhancers in the human genome (Hariprakash and Ferrari 2019). In addition, the initial choice of epigenomic marks used to map enhancers has a large impact on the final number and characteristics of defined enhancers (Zentner and Scacheri 2012). I showed that enhancer-gene association methods differed in the algorithmic details, but also in the way multiple parameters and information were used to define enhancers and promoters or quantify their activities. I further used this information to improve the reg2gene - a novel computational method that models *gene expression ~ enhancer activity.*

In **Chapter 4**, I explained the process of developing the reg2gene method. In short, reg2gene was built upon extensive data modeling and integration of the largest collection of epigenomics or transcriptomic data in humans at time - the Roadmap datasets (Roadmap Epigenomics Consortium et al. 2015) and its five sub-datasets: H3K4me1, H3K27ac, DNAme, DHS and RNA-Seq. reg2gene implements three correlation-based methods: Pearson, Spearman and distance correlation and executes the elastic net regression (Zou and Hastie 2005) and random forest (Breiman 2001) to account for the fact that multiple enhancers can act on a gene in a cooperative fashion (Reuter et al. 2015). Likewise, since an accurate enhancer definition represents a cornerstone upon which enhancer-gene models should be built (Hariprakash and Ferrari 2019), I performed modelling using a “consensus” enhancer definition. Each enhancer-gene pair (enhancers +/-1Mb around each TSS) was modelled a total of twenty times, and models were integrated by the majority voting approach and further improved by an ensemble voting with previously reported enhancer-gene associations (EGAs). I identified two sets of enhancer-gene

associations: a flexible set of ~230K EGAs reported in at least 2 publications and a stringent set of ~60 EGAs reported in three or more publications.

In **Chapter 5**, I analyzed sets of enhancer-gene associations in the light of multiple benchmarking datasets. The main idea behind this analysis was to test whether some of the computational methods predicted more accurate EGAs than other methods. As a benchmark datasets, I selected datasets that have been frequently used to benchmark computationally predicted EGAs (Gao et al. 2016, Cao et al. 2017, Hait et al. 2018): eQTLs and chromatin interactions. However, I showed that such benchmark datasets suffer from low reproducibility and doubted results of our and previous benchmarking procedures. I ran additional benchmarking with high-confidence *cis* enhancer-gene interactions assessed by cell-based CRISPR/Cas9 genetic screen (Gasperini et al. 2019) and I proposed an approach that searches for enhancer-gene "negative" associations. This approach enabled us to, for the first time, fully assess the performance of computational methods and assess all the elements of the confusion matrix and demonstrate that *stringentC* models have the highest PPV (positive predictive value) of 1.

In **Chapter 6**, I present results of SNP-to-gene annotation analysis performed using different sources of enhancer-gene associations. Specifically, I annotated risk polymorphisms from the GWAS Catalog, colorectal cancer (CRC) SNPs, rs104111210 and I identified that sets of annotated genes varied in their size. I show that, although the *stringentC* method could predict only 14 CRC genes, 65% of them were previously associated with CRC. I demonstrated that some of the well-known gene-CRC associations were missed by certain EGAs methods and present in other annotations, which indicated that each method could potentially add another piece of information necessary to improve our understanding of genetic susceptibility. Using benchmark datasets, I showed that novel SNP-CRC associations could be detected. Lastly, I identified examples of enhancer-based pleiotropy and novel gene-disease association.

7.2. Conclusions

The results presented in this thesis show that, even today, when a large number of methodological solutions has been proposed to map enhancers to their putative genes, we still do not have a systematic assessment of enhancer-gene associations in a genome-wide manner.

Data integration and availability of novel epigenomic datasets has been slowly improving our predictions of enhancer-gene associations, but we might never be able to identify a single “golden-standard” approach for discovering and documenting enhancer-gene interactions in a genome-wide manner.

Nonetheless, we should continue integrating information from various data sources to improve our understanding of gene regulation in health and disease.

Even in the light of aforementioned limitations, enhancer-gene associations represent an exciting resource for annotating risk SNPs to their target genes.

7.3. Future perspectives

The main bottleneck of computational models of enhancer-gene associations is the requirement for a large number of available cell types with comparable quality and resolution of data (Hariprakash and Ferrari 2019). Since this is an area of extensive scientific efforts - IHEC (The International Human Epigenome Consortium, Stunnenberg et al. 2016) and the shared effort of its nine members - ENCODE, Roadmap Epigenomics, BLUEPRINT, DEEP, Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC; Canada) together with the national epigenome projects from Japan, Korea, Singapore and Hong Kong recently set off to produce 1,000 reference epigenomes - we could soon expect major improvement in the number and quality of available datasets. The IHEC project is still ongoing, but it already represents the largest collection of tissues and cell types for which epigenomic profiling has been systematically done (Bae 2013; Albrecht et al. 2016). I would expect that reg2gene models would achieve better accuracy and precision if they would be trained using the IHEC datasets. In this modelling scenario, we would simply increase the number of training examples and reduce the hurdle of "large p, small n" problem (high-dimensional data with few examples). I would expect that, if not me, somebody will perform such analysis in the future.

I hypothesize that meta-analysis can be very useful in genomic studies as well (similar to its proven usefulness in the genome-wide association studies; Cantor et al. 2010). Meta-analysis is a statistical method that combines results of different studies, especially those with small sample

size or with conflicting results. It was often used to combine information from multiple GWAS and can increase the chances of finding true positives among the identified associations. In the case of IHEC epigenomes, it can be used to combine results of different consortium datasets to overcome their inherent differences and biases.

I believe that, based on current evidence, epigenome editing is the key way to improve our knowledge about enhancer-gene targeting. Genetic manipulation approaches can provide evidence for the importance of certain regulatory regions, which can be subsequently used to train computational algorithms. For example, site-specific epigenomic editing has been done by recruiting chromatin-modifying enzymes to specific loci using the CRISPR-Cas system: deactivated Cas9 - dCas9 (Jinek et al. 2012) was already used in a combination with a range of chromatin-modifying enzymes, general transcriptional activator or repressor proteins such as p300 (Hilton et al. 2015), LSD1 (Kearns et al. 2015), DNMT3A (Rivenbark et al. 2012; Siddique et al. 2013), KRAB (Fulco et al. 2016) to add or remove chromatin marks at the target. Recent experiments (Fulco et al. 2019; Gasperini et al. 2019) combine CRISPR-Cas perturbations with single-cell analyses (flow cytometry, expression profiling, etc.). Altogether, I hope that larger and more robust sets of enhancers-gene links will be accessible in the near future to train better computational models.

7.3. Outlook

The immediate application of computational modelling of enhancer-gene associations can be seen in human genetics: annotation of the risk-associated polymorphisms with genes they truly regulate represents a crucial step in understanding disease etiology of almost all human diseases. I presented certain examples of it in this thesis, but there is many more left to be explored.

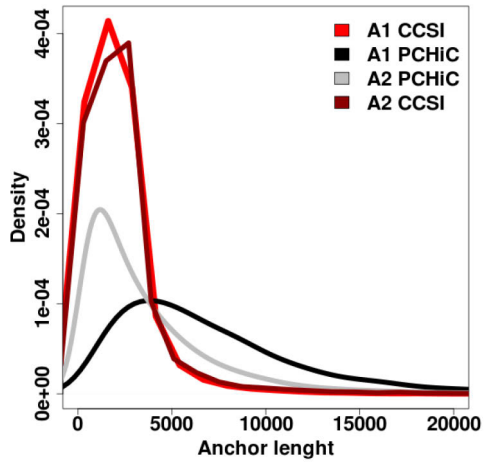
On the other hand, as a bioinformatician, I spent a lot of my working time visualizing ChIP-Seq signals for H3K4me1 and H3K27ac in the IGV Browser (Thorvaldsdóttir et al. 2013). Many of the visualized peaks seemed to be (function) enhancer regions. However, I could not discern which genes such regions regulate. With a database of EGAs, I can simply screen enhancers and potentially identify their targets.
In addition, reg2gene package can be simply used to support further modelling of similar biological questions that can further refine gene expression across cell types and developmental stages.

Lastly, many questions about EGAs remained unanswered. Which enhancers regulate two genes at the same time? Which cell-types have active which enhancers?

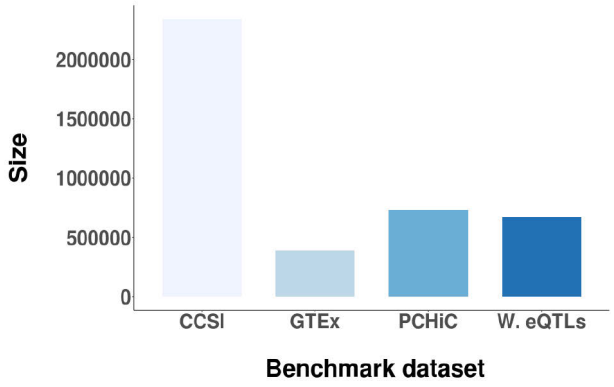
8

Supplement

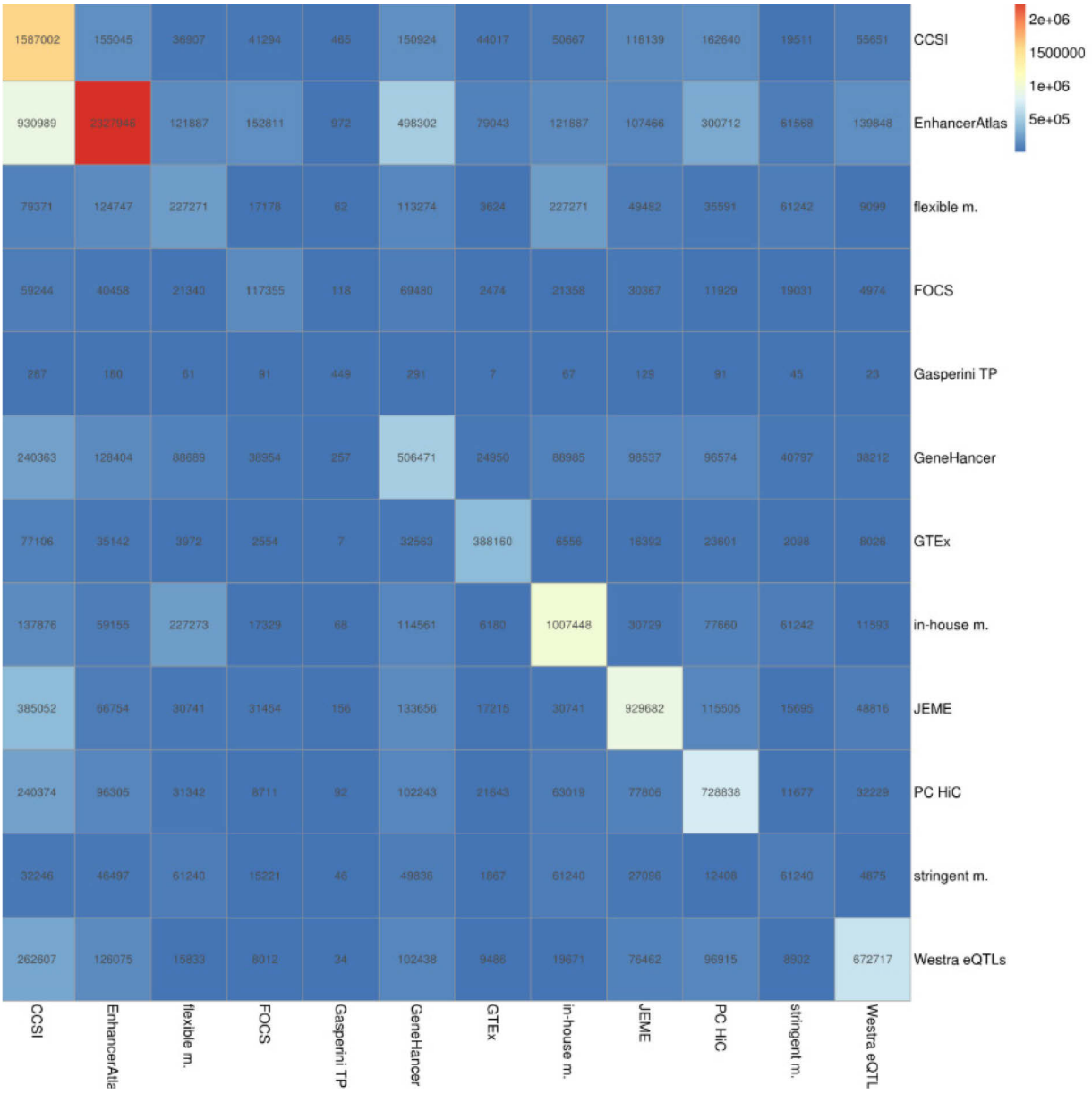
8.1 Supplementary figures



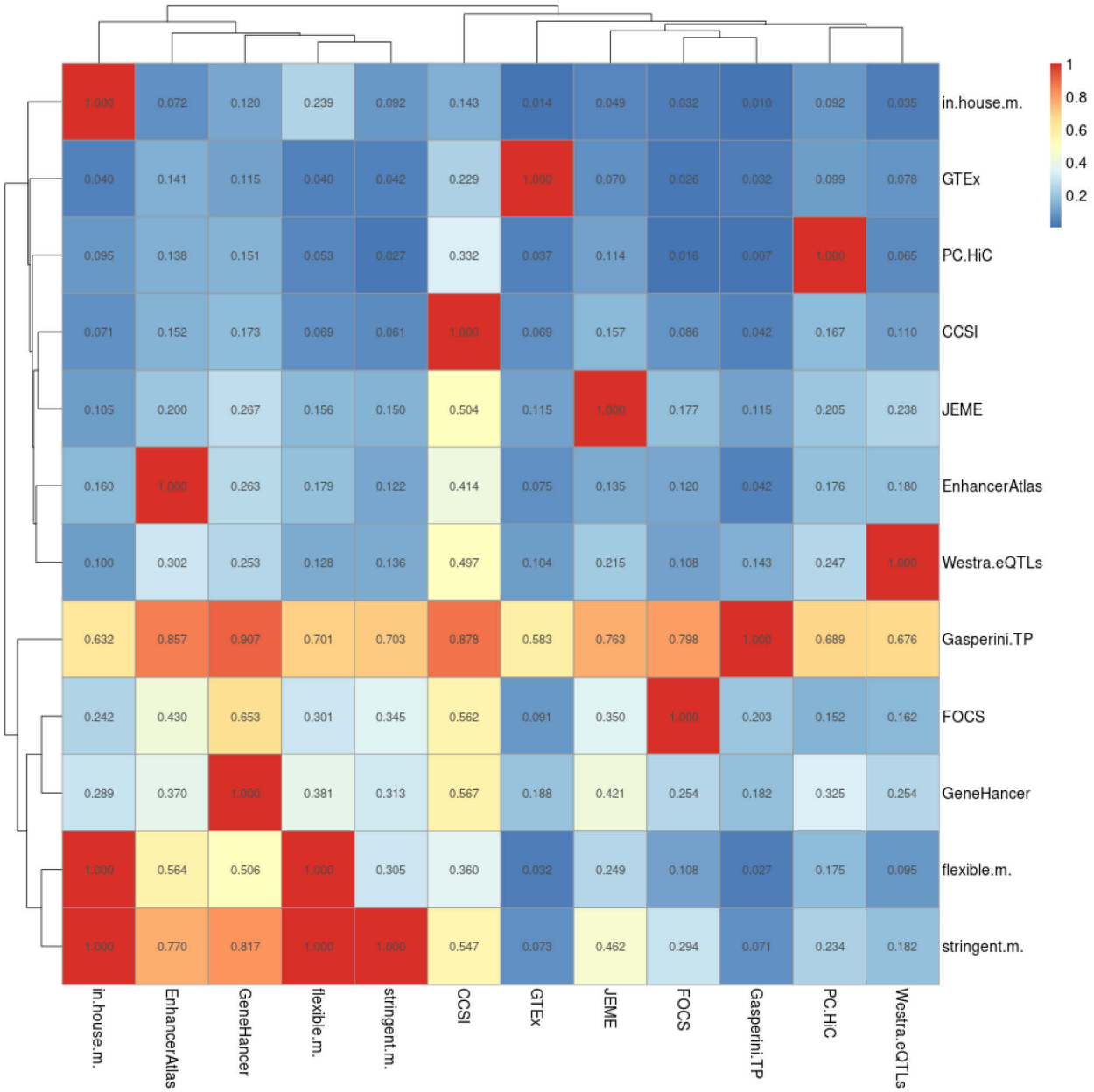
Supplementary Figure 1. Distribution of anchor sizes for the CCSI database and PC-HiC results form (Javierre et al., 2016). Since each reported chromatin, interacting pair consists of two anchors (A1 and A2), each of them was individually analyzed and distribution of their length was reported on the x-axis. On average, anchor1 (A1) and interacting anchor 2 (A2) had the same sizes, whereas anchors reported in the PC-HiC experiment varied in their size distribution.



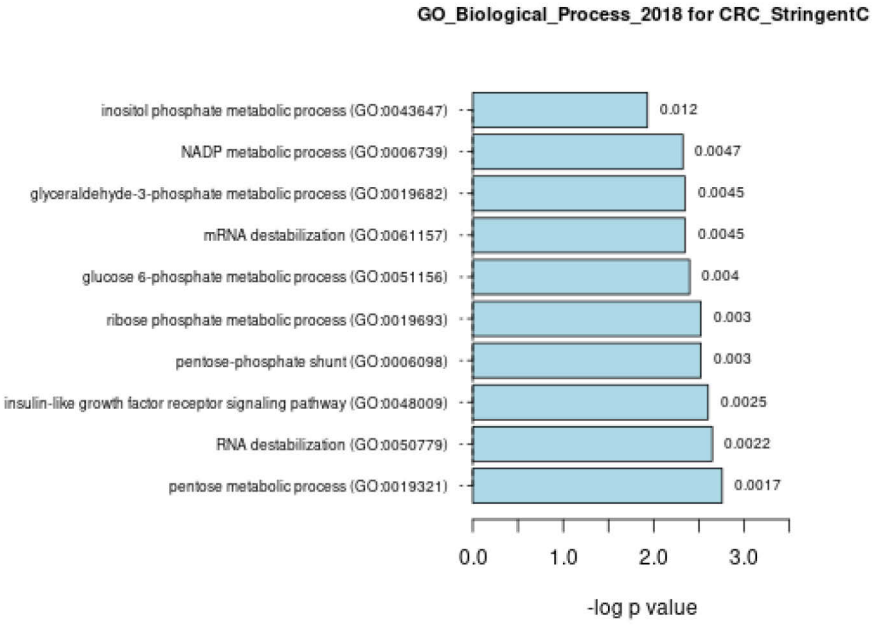
Supplementary Figure 2. Histogram of the number of interactions reported in each benchmark dataset: eQTLs from (Westra et al., 2013) and the GTEx database (GTEx Consortium et al., 2017) and PC-HiC experiments from (Javierre et al., 2016) and the CCSI database (Xie et al., 2016)

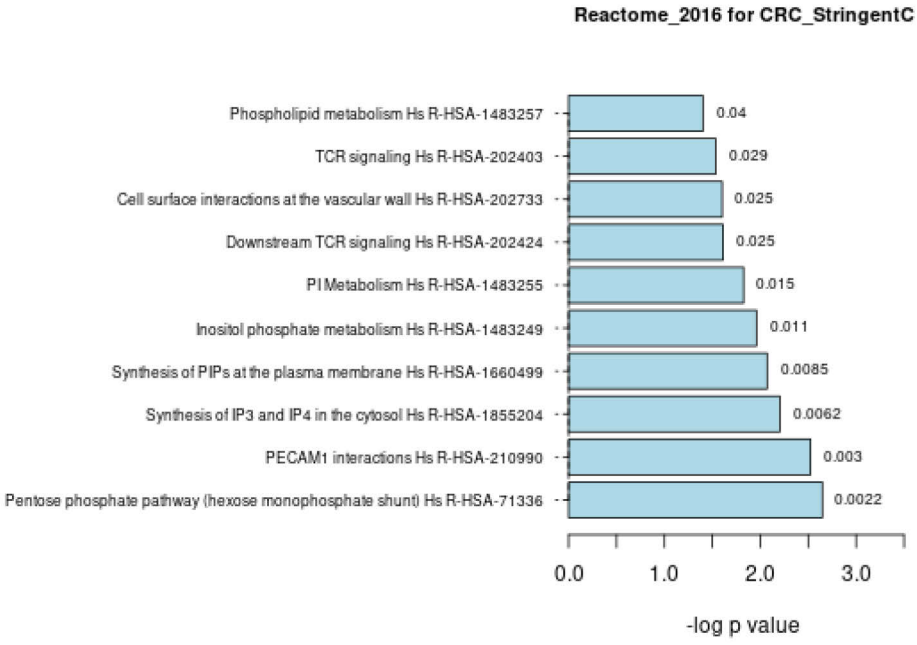


Supplementary Figure 3. Heatmap of the results of the benchmarking procedure for all analyzed datasets. The number of identified overlaps between two datasets was reported as the statistics. Here, we analyzed a total of 12 datasets: 7 sets of enhancer-gene associations (EnhancerAtlas (Gao et al., 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al., 2017), FOCS (Hait et al., 2018), reg2gene *inhouseM (in-house m.), stringentC (stringent m.) and flexibleC models (flexible m.);* two sources of eQTLs from (Westra et al., 2013, Westra eQTL) and the GTEx database (GTEx Consortium et al., 2017), GTEx), PC-HiC experiments from (Javierre et al., 2016, PC HiC) and the CCSI database (Xie et al., 2016); CCSI), and *cis*-regulatory interactions reported in (Gasperini et al., 2019). In rows, one can get information about the coverage of a given dataset, by other dataset. For example, out of 449 analyzed (Gasperini et al., 2019) reported interactions, 287 of them were covered by CCSI-reported chromatin interactions.

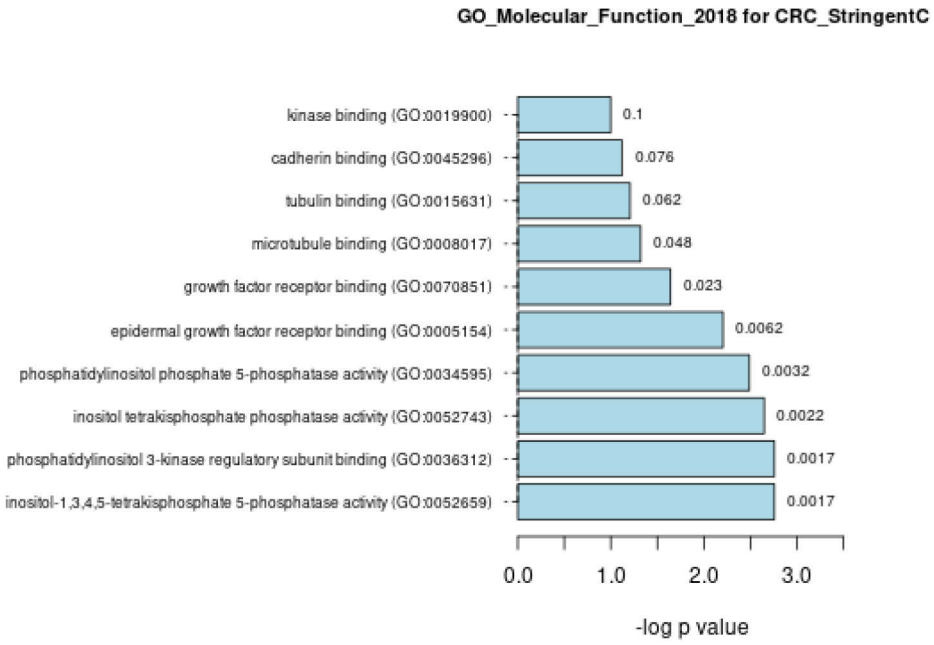


Supplementary Figure 4. Heatmap of the results of the benchmarking procedure for all analyzed datasets. The percentage of pairwise overlap between two datasets was reported as the statistics. We analyzed a total of 12 datasets: 7 sets of enhancer-gene associations (EnhancerAtlas (Gao et al., 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al., 2017), FOCS (Hait et al., 2018), reg2gene *inhouseM (in-house m.), stringentC (stringent m.) and flexibleC models (flexible m.);* two sources of eQTLs from (Westra et al., 2013, Westra eQTL) and the GTEx database (GTEx Consortium et al., 2017), GTEx), PC-HiC experiments from (Javierre et al., 2016, PC HiC) and the CCSI database (Xie et al., 2016); CCSI), and *cis*-regulatory interactions reported in (Gasperini et al., 2019). In rows, one can get information about the percentage of coverage of a given dataset, by other dataset. For example, out of 18.2% of *stringentC* models was covered by Westra eQTLs.

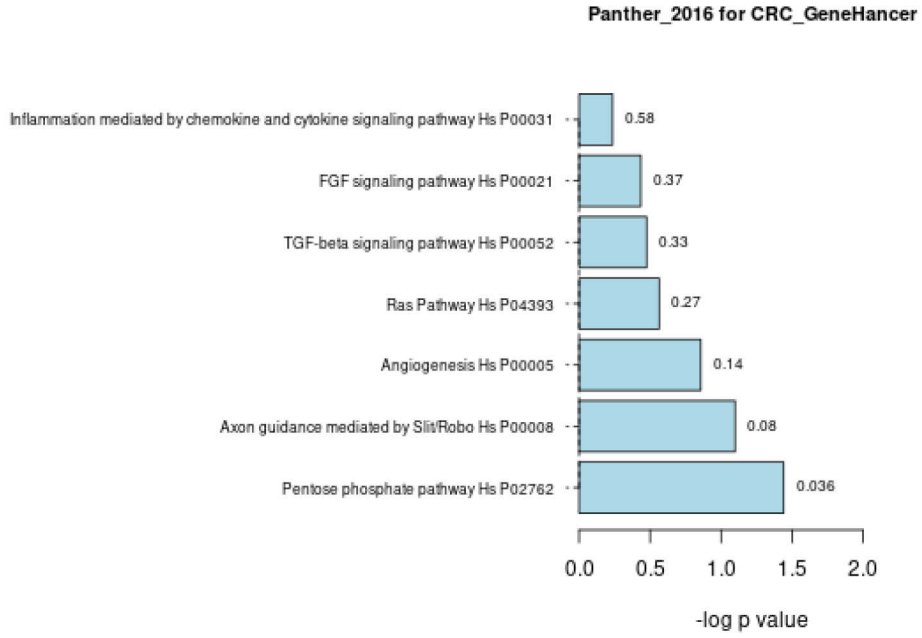

GO_Biological_Process_2018 for CRC_StringentC
inositol phosphate metabolic process (GO:0043647)
NADP metabolic process (GO:0006739)
glyceraldehyde-3-phosphate metabolic process (GO:0019682)
mRNA destabilization (GO:0061157)
glucose 6-phosphate metabolic process (GO:0051156)
ribose phosphate metabolic process (GO:0019693)
pentose-phosphate shunt (GO:0006098)
insulin-like growth factor receptor signaling pathway (GO:0048009)
RNA destabilization (GO:0050779)
pentose metabolic process (GO:0019321)
0.012
0.0047
0.0045
0.0045
0.004
0.003
0.003
0.0025
0.0022
0.0017
0.0
1.0
2.0
3.0
-log p value


Reactome_2016 for CRC_StringentC
Phospholipid metabolism Hs R-HSA-1483257
TCR signaling Hs R-HSA-202403
Cell surface interactions at the vascular wall Hs R-HSA-202733
Downstream TCR signaling Hs R-HSA-202424
PI Metabolism Hs R-HSA-1483255
Inositol phosphate metabolism Hs R-HSA-1483249
Synthesis of PIPs at the plasma membrane Hs R-HSA-1660499
Synthesis of IP3 and IP4 in the cytosol Hs R-HSA-1855204
PECAM1 interactions Hs R-HSA-210990
Pentose phosphate pathway (hexose monophosphate shunt) Hs R-HSA-71336
0.04
0.029
0.025
0.025
0.015
0.011
0.0085
0.0062
0.003
0.0022
0.0
1.0
2.0
3.0
-log p value

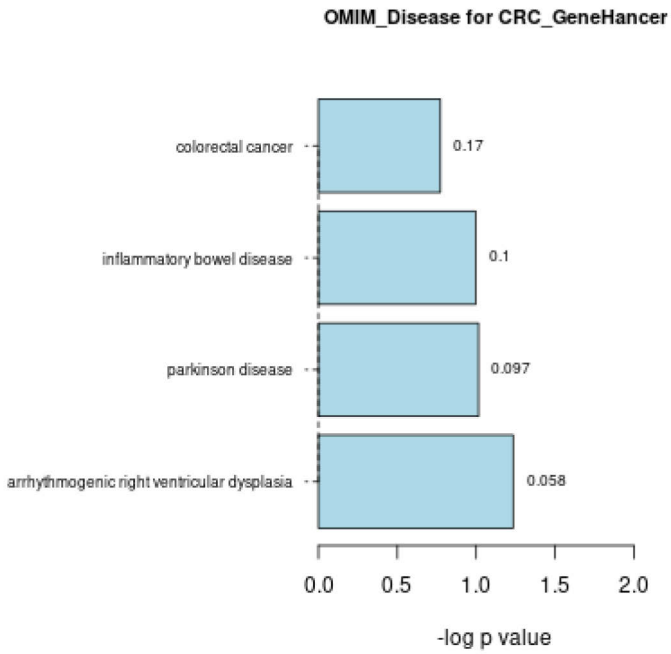
B.


GO_Molecular_Function_2018 for CRC_StringentC
kinase binding (GO:0019900)
cadherin binding (GO:0045296)
tubulin binding (GO:0015631)
microtubule binding (GO:0008017)
growth factor receptor binding (GO:0070851)
epidermal growth factor receptor binding (GO:0005154)
phosphatidylinositol phosphate 5-phosphatase activity (GO:0034595)
inositol tetrakisphosphate phosphatase activity (GO:0052743)
phosphatidylinositol 3-kinase regulatory subunit binding (GO:0036312)
inositol-1,3,4,5-tetrakisphosphate 5-phosphatase activity (GO:0052659)
0.1
0.076
0.062
0.048
0.023
0.0062
0.0032
0.0022
0.0017
0.0017
0.0
1.0
2.0
3.0
-log p value

C.

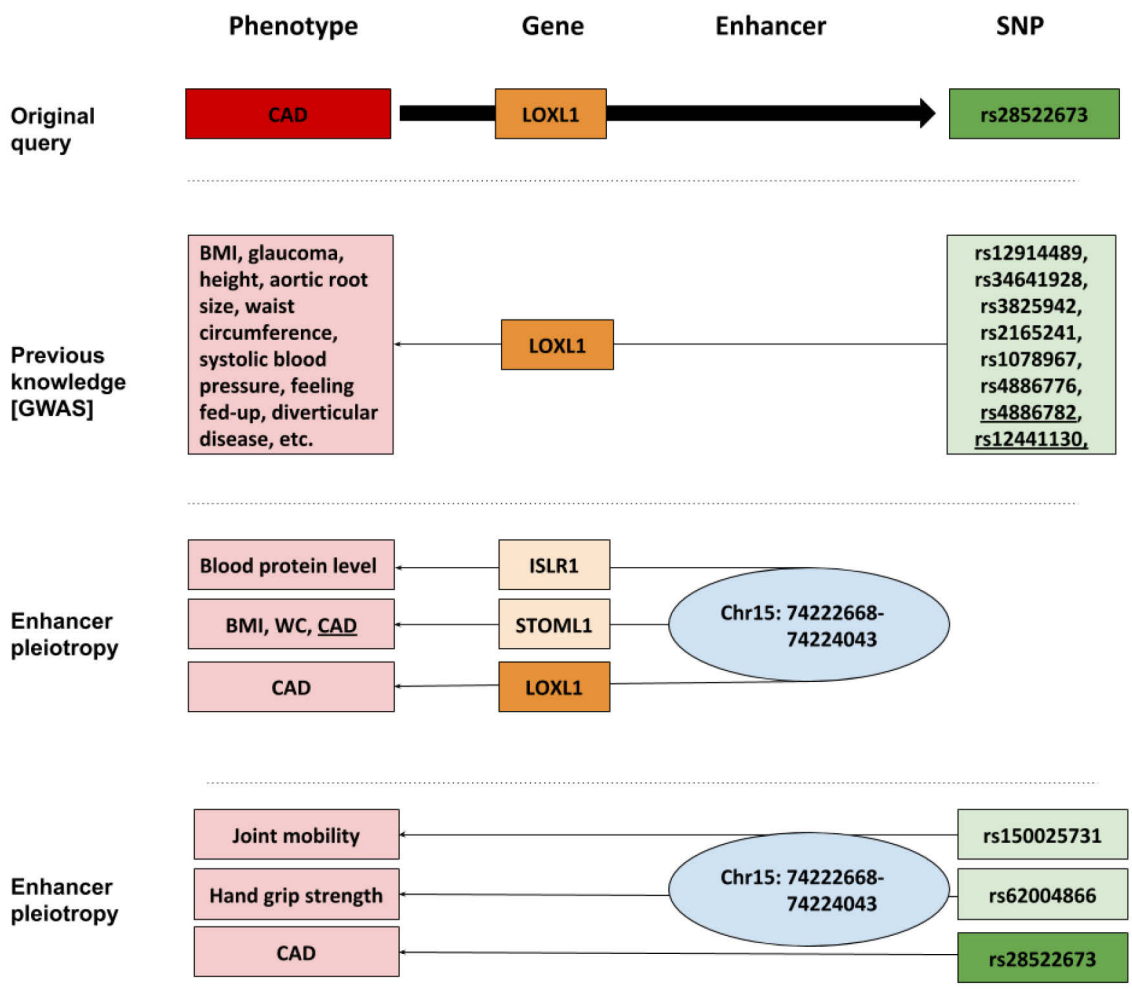

Panther_2016 for CRC_GeneHancer
Inflammation mediated by chemokine and cytokine signaling pathway Hs P00031
FGF signaling pathway Hs P00021
TGF-beta signaling pathway Hs P00052
Ras Pathway Hs P04393
Angiogenesis Hs P00005
Axon guidance mediated by Slit/Robo Hs P00008
Pentose phosphate pathway Hs P02762
0.58
0.37
0.33
0.27
0.14
0.08
0.036
0.0
0.5
1.0
1.5
2.0
-log p value

D.

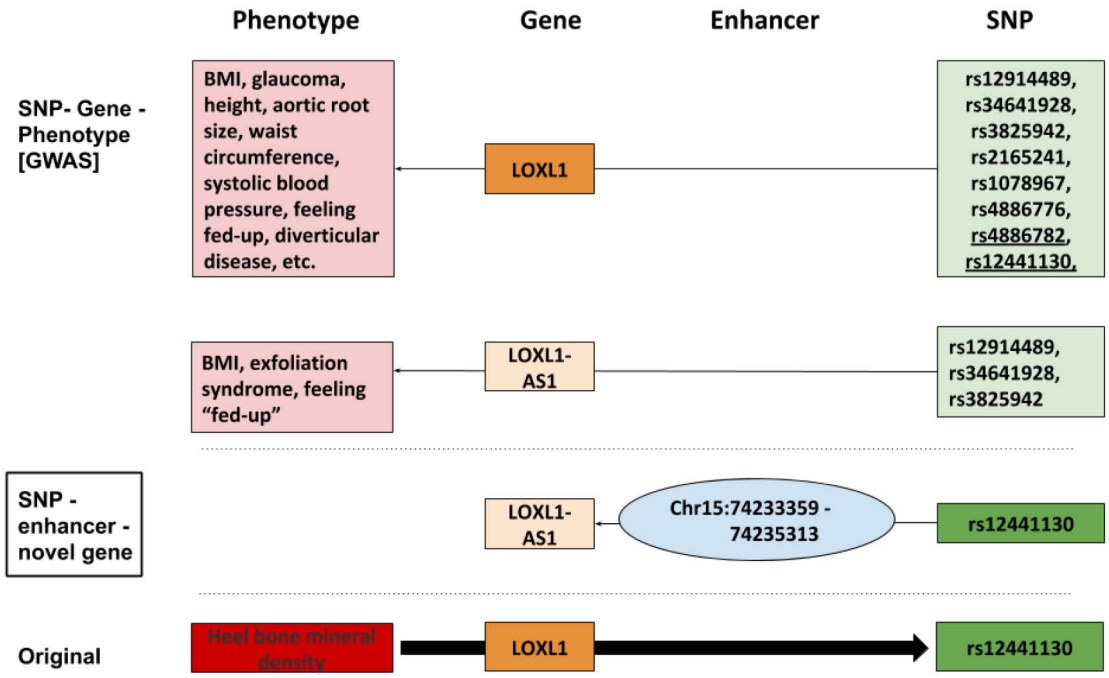


E.

Supplementary Figure 5. Results of the enrichment analysis for gene sets that were annotated to 312 CRC-associated SNPs using an overlap between SNP and enhancer regions from a set of enhancer-gene associations (EGAS). We performed gene enrichment analysis for: molecular function, cellular component, biological process, human phenotype ontology, OMIM disease, and KEGG, Reactome, and PANTHER pathways using enrichR R package (Kuleshov et al., 2016) and plotted TOP10 enrichment results (if applicable). A. Results of the enrichment analysis for biological processes for 14 CRC genes that were identified based on the overlap with *stringentC* EGAs. B. Results of the enrichment analysis for Reactome pathways for 14 CRC genes that were identified based on the overlap with *stringentC* EGAs. C. Results of the enrichment analysis for molecular function for 14 CRC genes that were identified based on the overlap with *stringentC* EGAs. D. Results of the enrichment analysis for Panther pathways for CRC genes identified using GeneHaner EGAs. E. Results of the enrichment analysis for OMIM diseases for CRC genes identified based on the SNP overlap with GeneHancer EGAs.



Supplementary Figure 6. A schematic representation of detecting plausible enhancer pleiotropy for the LOXL1 gene and its GWAS-associated SNPs. Original information is that rs28522673 is associated with cardiovascular disease via LOXL1 gene. Simply, by querying the GWAS Catalog we could supplement that information by identifying all SNPs and diseases associated with the LOXL1 gene. Using information about an overlap between rs28522673 and chr:7422268-74224043 enhancer regions, allowed us to detect putative enhancer pleiotropy. We identified an overlap between chr15:74222668-74224043 enhancer region and SNPs associated with different phenotypes (rs150025731 and rs62004866). In addition, we could identify an association of this enhancer region with non-LOXL1 genes: STOML1 and ISLR.



Supplementary Figure 7. A schematic representation of detecting plausible novel gene targets and extending previous knowledge for rs12441130. Original information was that rs12441130 was associated with heel bone mineral density via the LOXL1 gene. Simply by using information from the GWAS Catalog we could identify additional SNPs and diseases associated with the LOXL1 gene. On the other hand, we identifying an overlap between rs124411310 and chr15:74233359-7423513 enhancer region, thereby associating rs12441130 with novel gene target - the LOXL1-AS1 gene (and further identifying in the GWAS Catalog association of the LOXL1-AS1 gene with BMI, exfoliation syndrome, etc.

8.2. Supplementary tables

	N[enhancers]	N[genes]	N[EGA]	N[enh/gene]	N[gene/enh]	Enhancer length	N[repressed regions]	N[enhancer, expected]
EnhancerAtlas	1008243	51855	2327946	range: 1-608 [med: 24; mean: 44.9]	range: 1-52 [med: 2; mean: 2.3]	range: 11-3242861 [med: 1411; mean: 2852.4]	13672	NA
FOCS	103979	15131	117355	range: 1-69 [med: 6; mean: 7.8]	range: 1-15 [med: 1; mean: 1.1]	range: 3-16153 [med: 184; mean: 408.5]	274	92,603
GeneHancer	189301	27579	506471	range: 1-191 [med: 14; mean: 18.4]	range: 1-52 [med: 2; mean: 2.7]	range: 1-86253 [med: 1571; mean: 2412.7]	1121	284,834
JEME	291992	38817	929682	range: 1-127 [med: 20; mean: 24]	range: 1-80 [med: 2; mean: 3.2]	range: 3-2861 [med: 801; mean: 988.5]	272	489,581 Roadmap
InhouseM	184005	32151	1007448	range: 1-172 [med: 25; mean: 31.3]	range: 1-96 [med: 3; mean: 5.5]	range: 20-1999 [med: 711; mean: 821.7]	12	NA
FlexibleC	107489	24124	227271	range: 1-94 [med: 6; mean: 9.4]	range: 1-36 [med: 1; mean: 2.1]	range: 20-1999 [med: 770; mean: 858.1]	3	NA
StringentC	44559	14225	61240	range: 1-55 [med: 3; mean: 4.3]	range: 1-20 [med: 1; mean: 1.4]	range: 20-1999 [med: 844; mean: 898.7]	0	NA

Supplementary Table 1. Statistics behind four published methods: EnhancerAtlas (Gao et al., 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al., 2017), FOCS (Hait et al., 2018), and three in-house methods that predict enhancer gene associations: reg2gene *inhouseM, stringentC and flexibleC* models. For each method, we report numbers of enhancer and genes that interact in enhancer-gene associations. In addition, we calculated an average number of genes per enhancer, and enhancer per gene, as well as enhancer length. We report a number of enhancers that show full per-base overlap with regions predicted by ChromHMM as heterochromatin, repressed Polycomb, weak repressed Polycomb and quies. Lastly, we report the expected number of enhancers (number of enhancers reported in the individual publication).

Supplementary Table 2. Confusion matrix for enhancer-gene associations. Positives correspond to the *cis* enhancer-gene interactions from (Gasperini et al., 2019), whereas negative were defined by an in-house protocol that identified regions in the genome that were consistently inactive - across all 127 cell types in the

Roadmap dataset. TP correspond to *cis* enhancer-gene interactions reported in (Gasperini et al., 2019) that overlapped computationally predicted EGAs, whereas FN are *is* enhancer-gene interactions reported in (Gasperini et al., 2019) that were not predicted. FP are computationally predicted EGAs whose enhancers were predicted to be "inactive" by the ChromHMM algorithm across all 127 cell types reported in the Roadmap dataset (details in Methods section). TN are EGAs whose enhancers were predicted to be "inactive" by the ChromHMM algorithm across all 127 cell types reported in the Roadmap dataset and their association was not identified. TP=true positive, FP= false positive, TN= true negative, FN= false negative.

	TP [EGA]	FN [EGA]	FP [EGA]	TN [EGA]	PPV
stringentC	45	19	0	3,147	100%
flexibleC	61	26	4	3,143	93.4%
InhouseM	67	39	47	3,100	58.8%
FOCS	91	23	293	NA	23.7%
JEME	129	40	671	NA	16.1%
EnhancerAtlas	180	30	35,618	NA	0.5%
GeneHancer	291	30	1,635	NA	15.1%

Supplementary Table 3. General statistics behind the GWAS Catalog annotated SNPs across all seven EGA methods used to annotate SNPs. Number of SNPs annotated to enhancers are reported along with corresponding enhancer-gene annotations and underlying genes.

	SNPs	Genes	Entries	Enhancers
EnhancerAtlas	27769	28988	388702	70718
FlexibleC	3285	5198	14857	3030
FOCS	1651	1422	3535	1597
GeneHancer	10487	13287	79680	8601
InhouseM	4915	15013	57151	4569
JEME	6992	13595	45761	6524
StringentC	1601	1769	4756	1472

Supplementary Table 4. A list of genes that were annotated to 312 colorectal cancer SNPs from the GWAS Catalog using 7 sources of enhancer-gene associations: EnhancerAtlas (Gao et al., 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al., 2017), FOCS (Hait et al., 2018), and reg2gene *inhouseM* (IHM)*, stringentC* (Str) *and flexibleC* (Flex) models.

Method	Genes
EA	EIF1B, ZNF621, CTNNB1, ZNF619, RPL14, LRIG1, SLC25A26, RNU6-787P, COX15, BLOC1S2, CWF19L1, ENTPD7, SNORA12, ERLIN1, ARFGEF2, PREX1, DCBLD1, NUS1, PPME1, RPS3, SPCS2, GDPD5, MRPL48, RP11-707G14.1, AP000560.1, BMP4, CDKN3, SCG5, SMAD7, RP11-484L8.1, TRMT6, BMP2, CRLS1, MCM8, ADRM1, YTHDF1, MIA3, RP11-452F19.3, ZMIZ1, GATA3, TCF7L2, Y_RNA, MBNL1, ACSL5, ZDHHC6, VTI1A, RP11-57H14.3, RP11-57H14.4, C12orf4, PITX1, C5orf24, CXCL14, CAMLG, TXNDC15, MYBL2, CENPN, RNU6-925P, MTHFSD, RN7SL381P, BAALC, RP11-318M2.3, DHX35, PPP1R16B, SNORA71D, SNORA71A, RALGAPB, RP4-616B8.4, EVC2, LYAR, MBOAT7, SSC5D, TSEN34, RDH13, TMEM150B, DNAAF3, SRP54, BAZ1A, TAF1A, DUSP10, RP11-191N8.2, RP11-815M8.1, ACAT2, RNU4-25P, MYC, RHPN2, CEP89, ANKRD27, PEPD, NUDT19, NUTF2, RNU6-359P, RNU6-22P, ZFP90, CDH1, RP11-311C24.1, NFAT5, CDH3, SLC12A4, CTC-479C5.10, CDKN1A, FGD2, ETV7, SRSF3, KCTD20, ARMC12, CPNE5, SRPK1, RNU1-88P, TBX3, AASDHPPT, MSANTD4, RP11-408N14.1, MXD1, ZC3H7A, LITAF, RSL1D1, RP11-485G7.5, CTNNAL1, TMEM245, ZNF98, ZNF493, ZNF208, RNGTT, SCAF8, DGAT2, SERPINH1, SNORD15B, ARRB1, CTD-2530H12.1, MON1B, WWOX, FBXO9, ICK, RP11-79N23.1, NFIB, PREPL, AC009236.2, SSR1, PAPD7, GALNT10, GRIA1, NANS, SEC61B, TGFBR1, GALNT12, USPL1, B3GALTL, HSPH1, MEDAG, CDC42, LINC00339, NBPF3, WNT4, LNPEP, ERAP1, RP11-473I1.9, RP11-473I1.10, FOXF1, FOXL1, FOXC2, EMC8, RAD21, UTP23, AARD, EIF3H, KLF13, NABP1, GOPC, GSDMC, FAM49B, JKAMP, C14orf37, DACT1, PSMA3, DAAM1, TIMM9, EXTL3, DCTN6, DUSP4, RNA5SP260, LEPROTL1, INTS9, SLCO3A1, IGFBP3, LONRF1, LAMC1, RNU6-41P, DHX9, BICC1, PHYHIPL, CD9, TNFRSF1A, LTBR, MRPL51, GAPDH, CDCA3, PLEKHG6, ZNF384, C1S, LPAR5, SCNN1A, PTPN6, MLF2, PHB2, U47924.27, RP1-102E24.8, NCAPD2, EMG1, GLOD4, SLC43A2, RNMTL1, TIMM22, RP11-676J12.6, NXN, LARP4, COX14, FERMT1, NXPE3, ZBTB11, FRS3, TFEB, TRERF1, CCND3, PGC, MDFI, OARD1, RNU6-761P, TOMM6, RP11-298J23.5, TREM1, MYNN, TUBB, PPP1R10, RPP21, TRIM26, ZNRD1, AGPAT1, BRD2, PSMB9, PSMB8, TAP2, HLA-DQA2, C6orf48, HLA-DQB1, MSH5, HLA-DQB2, SNTA1, FRY, N4BP2L2, RALY, RP5-1125A11.1, IL6ST, MIER3, MAP3K1, MRGPRF, WNT1, KMT2D, SPATS2, TUBA1C, TROAP, EIF4B, KRT8, KRT18, ZFP36L1, BCL2L14, APOLD1, DDX47, SAMD3, TMEM200A, TRIT1, KCNS3, GEN1, RDH14, YEATS4, RAP1B, NUP107, FRS2, TSPAN11, SNORA75, CAPRIN2, FAM60A, LINC00941, SLC20A2, FOXQ1, KIAA0020, C1orf198, COG2, LIAS, SMIM14, UBE2K, TLR10, TLR1, UGDH, RFC1, TMEM156, TLR6, FASTKD3, BUB3, C5orf22, DROSHA, KIAA0355, ZNF181, GPI, TMTC1, TNFSF15, PRKG2, AGAP1, OSBP2, SLC35E4, EIF4ENIF1, ZNF542, ZNF460, AAMP, PNKD, TMBIM1, CYP27A1, ARPC2, VIL1, BCS1L, STK36, USP37, RP11-378A13.1, FAM83B, RP11-524H19.2, HSPA8, CLMP, GRAMD1B, PTPRM, NDUFV2, PUM1, SDC3, SNORD103B, FABP3, GADD45A, SERBP1, CXCR4, R3HDM1, DARS, FAM193A, ADD1, NOP14, SH3BP2, NEDD9, PAK1IP1, ADTRP, INSIG1, RBM33, RANBP6, UHRF2, OSTF1, NMRK1, PTGR1, DNAJC25, RNA5SP295, PTBP3, SUSD1, RN7SL57P, ATP5O, SLC5A3, DNAJC28, SMIM11, DAB2, KIAA2018, MIR568, RASGRF2, SSBP2, ASB6, CCBL1, CRAT, TMSB4XP4, PPP2R4, IER5L, ZDHHC12, COQ4, DOLPP1, GOLGA2, TOR1B, NTMT1, NUP188, DOLK, LRRC8A, CERCAM, GMDS, RP11-434C1.1, ETV6, RP11-180M15.3
Flex	DUSP10, MYC, CASC8, RP11-514D23.3, MTHFSD, SLC7A9, RHPN2, PEPD, HOXB9, RP11-678G14.2, RP11-678G14.3, ZNF208, RPS3, SERPINH1, UVRAG, GDPD5, MAP6, NANS, COL15A1, RP11-191N8.2, RP11-400N13.3, KLF13, MTMR10, DCBLD1, ROS1, RP1-92C8.3, GOPC, ATP5C1, TAF3, RP11-379F12.4, MTG2, ADRM1, MRGBP, GID8, DIDO1, DVL1, MXRA8, FLJ31104, AC008940.1, MAP3K1, IL6ST, FGFR2, FAM83B, PPP2R4, PRRX2, TOR1B, ZDHHC12, PTGES
FOCS	BMP2, RHPN2, GPATCH1, TET2, HOXB-AS1, HOXB2, HOXB3, HOXB6, HOXB9, SERPINH1, DCBLD1, TFEB, DVL1, MXRA8, FAM83B, SH3GLB2
GH	PREX1, ARFGEF2, DDX27, RP1-92C8.3, DCBLD1, COLCA2, C11orf53, COLCA1, BMP4, SMAD7, RPL17, LIPG, DUSP10, RP11-481G8.2, ZMIZ1-AS1, TCF7L2, CASC8, MYC, RP11-568J23.5, RP11-514D23.3, MTHFSD, FADS2, FADS1, FADS3, RAB3IL1, DKFZP434K028, MYRF, INCENP, BEST1, POLR2G, CD83, LAMA5, CABLES2, LAMA5-AS1, GPATCH1, CEP89, RHPN2, FAAP24, ZNF507, ANKRD27, PDCD5, NUDT19, ZFP90, TANGO6, CDH1, CDKN1A, SRSF3, FGD2, HOXB2, HOXB-AS1, HOXB1, RP11-678G14.3, RP11-678G14.2, ZNF429, ZNF708, ZNF100, LINC00664, FBP1, ZNF169, C9orf3, GDPD5, RPS3, UVRAG, SERPINH1, EFHC1, ELOVL5, FBXO9, AC009236.2, RP11-69L16.4, COL15A1, GALNT12, RP11-58A18.1, GINS2, CERS5, DIP2B, MTMR10, KLF13, CCND2-AS2, PARP11, CCND2, C12orf4, TIGAR, RBM19, TBX3, MORC1-AS1, MORC1, RP11-316M21.7, LINC01475, SLC25A28, DNMBP, NKX2-3, CD9, RP11-676J12.6, NXN, WDR88, AC092301.3, DMWD, SYMPK, DMPK, RSPH6A, RP11-298J23.5, PGC, TFEB, RP11-25C19.3, MYH3, ADPRM, RP11-157P1.4, DIDO1, TAF4, RPS21, MRGBP, N4BP2L1, RP5-1125A11.4, MAP1LC3A, RALY, AHCY, RALY-AS1, ANKRD65, TAS1R3, DVL1, AC008940.1, SETD9, C5orf67, FGFR2, ETV6, BCL2L14, RP11-102N11.1, TMEM200A, SAMD3, BMP2, MCM8, RP11-420K8.1, HSPE1P8, LYZ, FRS2, CCT2, YEATS4, AC011515.2, FCAR, LILRB4, LAIR1, LILRB1, KIR3DX1, LILRB2, KIR2DL3, LILRA5, LILRB3, LILRA1, LILRA2, PPP1R12C, LENG8, LAIR2, LILRA6, KIR3DL3, LINC00941, CAPRIN2, ANK1, AP3M2, GPAT4, RPL9, SMIM14, AC006003.3, PTPRN2, PLD1, TMEM212, RAB31, SLC11A1, CTDSP1, ARPC2, PNKD, TMBIM1, CATIP-AS1, GPBAR1, USP37, CXCR2, ZNF142, CXCR4, UBXN4, MCM6, RNF4, FAM193A, RP11-276E15.4, SUSD1, HSDL2, PRPF4, DNAJC25, SNX30, PTBP3, HSPE1P28, CARD6, CRAT, NTMT1, C9orf106, PRRX2, LIPN, LIPF, LIPJ, RNLS, LIPM, ANKRD22

Method	Genes
IHM	CTD-2562J17.9, UCP2, DUSP10, RP11-191N8.2, RP11-815M8.1, RP11-400N13.1, RP11-358H9.1, RP11-506M13.3, RP11-17G2.1, RP11-31E13.2, NUTM2B-AS1, CTSLP6, RP11-57H14.3, NHLRC2, RP11-324O2.3, ZDHHC6, RP11-25C19.3, MYC, CASC8, LINC00311, RP11-680G10.1, LINC01082, FOXF1, RP11-805I24.4, RP11-805I24.3, RP11-514D23.3, RP11-514D23.2, RP11-158I3.1, FENDRR, MTHFSD, CTD-2085J24.3, CTD-2540B15.9, CEBPG, SLC7A9, RHPN2, CTD-2540B15.12, PEPD, RP11-67A1.2, NFATC3, VPS4A, DPEP3, SP2, CDK5RAP3, RP11-433M22.2, HOXB-AS3, PRAC2, SCRN2, HOXB8, HOXB9, SUMO2P17, RP11-420K14.1, ZNF626, RP11-678G14.2, RP11-678G14.3, RP11-678G14.4, ZNF208, ZNF98, RNF169, RP11-147I3.1, NEU3, RPS3, SERPINH1, CTD-2530H12.4, CTD-2530H12.2, UVRAG, XRRA1, CTD-2562J17.4, RP11-939C17.4, GDPD5, CTD-2530H12.8, CTD-2530H12.7, MAP6, RP11-535A19.2, CTD-2011F17.2, PRKRIR, NANS, RP11-92C4.4, COL15A1, CORO2A, MIA3, C1orf140, RP11-400N13.3, AIDA, CTD-3092A11.1, AC026150.6, AC026150.8, GOLGA8Q, GOLGA8H, RP11-16E12.1, KLF13, RP11-717I24.1, DNM1P31, DNM1P32, RP11-932O9.7, MTMR10, RP11-16E12.2, GOLGA8K, ULK4P1, RP13-395E19.3, FAM26E, FAM26D, KPNA5, RAP1BP3, DCBLD1, TRAPPC3L, ROS1, RP1-92C8.3, RP1-179P9.3, GOPC, ATP5C1, TAF3, RP11-379F12.4, RP11-379F12.3, GATA3-AS1, RP5-1029F21.2, MTG2, ADRM1, LAMA5-AS1, MRGBP, GID8, RP11-429E11.3, DIDO1, RP11-206L10.9, CPTP, VWA1, ATAD3A, RP1-283E3.4, RP11-206L10.8, NOC2L, HES4, SDF4, UBE2J2, DVL1, MXRA8, AURKAIP1, TMEM240, CFAP74, FLJ31104, CTD-2227I18.1, AC008940.1, MAP3K1, AC016644.1, IL6ST, ANKRD55, RP11-155L15.1, AC008937.2, FGFR2, BCDIN3D, RAB3IP, GINS4, IKBKB, RP11-360L9.7, RP11-360L9.4, DKK4, MTCL1, GACAT2, RP11-856M7.6, FAM83B, SMIM13, RP3-510L9.1, HIVEP1, ELOVL2, URM1, RP11-216B9.8, PPP2R4, RP11-344B5.4, RP11-344B5.3, LINC01503, PRRX2, TOR1B, USP20, RP11-409K20.6, ODF2-AS1, RP11-216B9.9, ZDHHC12, PTGES, FNBP1
JEME	NKX2-3, GOT1, COX15, SLC25A28, ENTPD7, UTP23, EIF3H, BMP2, PLCH1, CCND2, C12orf5, TOX2, CALB1, GOLGA8A, ACTC1, AVEN, LPCAT4, GJD2, NOP10, NUTM1, ZNF770, FAM83D, DHX35, PPP1R16B, ACTR5, AMPH, FAM183B, VPS41, POU6F2, NLRP2, GP6, TNNI3, EPS8L1, RDH13, NLRP7, PPP1R12C, SLC7A9, C19orf40, GPATCH1, RHPN2, WDR88, LRP3, CEP89, SLC7A10, CCDC28A, ECT2L, NHSL1, MGAT5B, MXRA7, SEC14L1, TSHZ1, RNGTT, AL079342.1, TDP2, ACOT13, C6orf62, KIAA0319, GPLD1, ALDH5A1, GMNN, SERPINH1, GDPD5, RPS3, KLHL35, ARRB1, MAP6, RP11-597K23.2, MEX3B, MED10, GRIA1, C16orf72, KLF13, DACT1, PCBD2, PITX1, CTC-349C3.1, CATSPER3, CLPTM1L, SLC12A7, TERT, SLC6A18, SLC6A19, TUBA1B, C1QL4, SPATS2, TUBA1A, TUBA1C, PRPH, LMBR1L, TROAP, KRT18, KRT8, EIF4B, KRT78, KRT4, SPRYD3, KRT3, KRT79, KRT76, TENC1, IGFBP6, ANK1, NKX6-3, AL033381.1, MYO16, ZAP70, ACTR1B, COX5B, PATZ1, DRG1, SFI1, AC005003.1, USP29, ZNF264, PEG3, ZIM3, DUXA, RAB12, SOGA2, GADD45A, GNG12, ADTRP, TMEM170B, CKMT2, ZCCHC9, IER5L, CRAT, PPP2R4, DOLPP1, FAM73B, SH3GLB2, PHYHD1, DOLK, CIZ1, NUP188, C9orf106, STAMBPL1, ANKRD22, LIPM, LIPK, FAS, ACTA2, LIPN, LIPF, LIPJ, RNLS
Str	DUSP10, MTHFSD, RHPN2, RPS3, SERPINH1, GDPD5, KLF13, DCBLD1, GOPC, DIDO1, DVL1, FAM83B, PPP2R4, TOR1B

Supplementary Table 5. General statistics behind annotation results for rs10411210 SNPs across all seven EGA methods used to annotate it. Number of enhancers, genes and enhancer-gene associations (EGAs) annotated to rs10411210 are indicated. In this case, results were calculated for benchmark datasets as well: two sources of eQTLs from (Westra et al., 2013, Westra eQTL) and the GTEx database (GTEx Consortium et al., 2017), GTEx), PC-HiC experiments from (Javierre et al., 2016, PC HiC) and the CCSI database (Xie et al., 2016); CCSI), IHM = *inhouseM;* eQTLW = Westra eQTLs.

	Publication	Enhancer	Gene	EGA
1	EnhancerAtlas	11	5	18
2	eQTLW	1	1	1
3	FlexibleC	2	3	5
4	FOCS	2	2	3
5	GeneHancer	1	8	8
6	GTEx	1	1	1
7	IHM	2	7	9
8	JEME	1	8	8
9	StringentC	2	1	2

Supplementary Table 6. Enhancer-gene pairs annotated to rs10411210 using information from nine data sources. Described in the figure above.

Enhancer	EnhancerWidth	Gene	Method
chr19:33532150-33532590	441	PEPD	EA
chr19:33532220-33532620	401	PEPD	EA
chr19:33530820-33535150	4331	PEPD	EA
chr19:33529860-33535150	5291	NUDT19	EA
chr19:33532150-33532590	441	RHPN2	EA
chr19:33527120-33535150	8031	RHPN2	EA
chr19:33531670-33532970	1301	RHPN2	EA
chr19:33530820-33535150	4331	RHPN2	EA
chr19:33530660-33535150	4491	RHPN2	EA
chr19:33529640-33534930	5291	RHPN2	EA
chr19:33528480-33533700	5221	RHPN2	EA
chr19:33529860-33535150	5291	RHPN2	EA
chr19:33531670-33532970	1301	CEP89	EA
chr19:33531290-33533640	2351	CEP89	EA
chr19:33532220-33532620	401	ANKRD27	EA
chr19:33531670-33532970	1301	ANKRD27	EA
chr19:33532150-33532590	441	ANKRD27	EA
chr19:33531840-33532950	1111	ANKRD27	EA
chr19:33531704-33533702	1999	SLC7A9	Flex
chr19:33531704-33533702	1999	RHPN2	Flex
chr19:33531704-33533702	1999	PEPD	Flex
chr19:33532126-33532564	439	RHPN2	Flex
chr19:33532126-33532564	439	PEPD	Flex
chr19:33532238-33532582	345	RHPN2	FOCS
chr19:33532238-33532582	345	GPATCH1	FOCS
chr19:33531700-33533702	2003	RHPN2	FOCS
chr19:33529636-33535923	6288	GPATCH1	GH
chr19:33529636-33535923	6288	CEP89	GH
chr19:33529636-33535923	6288	RHPN2	GH
chr19:33529636-33535923	6288	FAAP24	GH
chr19:33529636-33535923	6288	ZNF507	GH
chr19:33529636-33535923	6288	ANKRD27	GH
chr19:33529636-33535923	6288	PDCD5	GH
chr19:33529636-33535923	6288	NUDT19	GH
chr19:33532300-33532300	1	RHPN2	GTEx
chr19:33531704-33533702	1999	CTD-2085J24.3	IHM
chr19:33531704-33533702	1999	CTD-2540B15.9	IHM
chr19:33531704-33533702	1999	CEBPG	IHM
chr19:33531704-33533702	1999	SLC7A9	IHM
chr19:33531704-33533702	1999	RHPN2	IHM
chr19:33531704-33533702	1999	CTD-2540B15.12	IHM
chr19:33531704-33533702	1999	PEPD	IHM
chr19:33532126-33532564	439	RHPN2	IHM
chr19:33532126-33532564	439	PEPD	IHM
chr19:33532238-33532582	345	SLC7A9	JEME
chr19:33532238-33532582	345	C19orf40	JEME
chr19:33532238-33532582	345	GPATCH1	JEME
chr19:33532238-33532582	345	RHPN2	JEME
chr19:33532238-33532582	345	WDR88	JEME
chr19:33532238-33532582	345	LRP3	JEME
chr19:33532238-33532582	345	CEP89	JEME
chr19:33532238-33532582	345	SLC7A10	JEME
chr19:33531704-33533702	1999	RHPN2	Str
chr19:33532126-33532564	439	RHPN2	Str
chr19:33532300-33532300	1	LRP3	eQTLW

Supplementary Table 7. SNP and genes that were identified to be in LD (R2=0.8) with CRC-associated index SNP rs10411210

SNPs	Genes
rs10404631	RHPN2, GPATCH1, ZNF507, CEP89, WDR88
rs10411210	PEPD, NUDT19, RHPN2, CEP89, ANKRD27, SLC7A9, GPATCH1, FAAP24 , ZNF507, PDCD5, CTD-2085J24.3, CTD-2540B15.9, CEBPG, CTD-25 40B15.12, C19orf40, WDR88, LRP3, SLC7A10
rs10424333	RHPN2, GPATCH1, ZNF507, CEP89, WDR88
rs11880141	RHPN2
rs11881367	RHPN2, GPATCH1, PDCD5, LRP3, SLC7A9, ANKRD27
rs12459751	GPATCH1, LRP3, RHPN2, KCTD15, CTD-2540B15.7, ANKRD27, ZNF507 , CEP89, PDCD5, FAAP24, NUDT19
rs28363937	GPATCH1, RHPN2, KCTD15, CTD-2540B15.7, LRP3, AC007773.2, ANK RD27, ZNF507, CEP89, PDCD5, FAAP24, NUDT19
rs28403377	GPATCH1, RHPN2, KCTD15, CTD-2540B15.7, LRP3, AC007773.2, ANK RD27, ZNF507, CEP89, PDCD5, FAAP24, NUDT19
rs28505093	GPATCH1, RHPN2, CEP89, LRP3, C19orf40, WDR88
rs60507951	PEPD, GPATCH1, RHPN2, CTD-2085J24.4, CTD-2540B15.7, LRP3, AN KRD27, RN7SKP22, CEP89, ZNF507, FAAP24, DPY19L3, WDR88
rs7247582	GPATCH1, RHPN2, KCTD15, CTD-2540B15.7, LRP3, AC007773.2, ANK RD27, ZNF507, CEP89, PDCD5, FAAP24, NUDT19
rs7249860	RHPN2, GPATCH1, PDCD5, LRP3, SLC7A9, ANKRD27
rs7250288	GPATCH1, CEP89, LRP3, C19orf40, RHPN2, WDR88
rs7255601	RHPN2, GPATCH1, PDCD5, LRP3, ANKRD27
rs7258173	NUDT19, RHPN2, GPATCH1, PDCD5, LRP3, C19orf40, CEP89, ANKRD2 7, ZNF507, WDR88
rs73585909	GPATCH1, RHPN2, KCTD15, CTD-2540B15.7, LRP3, ANKRD27, ZNF507 , CEP89, PDCD5, FAAP24, NUDT19
rs73585910	GPATCH1, RHPN2, KCTD15, CTD-2540B15.7, LRP3, ANKRD27, ZNF507 , CEP89, PDCD5, FAAP24, NUDT19

Supplementary Table 8. Statistics behind an overlap across three CRC benchmark gene sets and CRC-annotated gene sets. Benchmark CRC gene sets correspond to: 1) DisGeNET - 1,676 CRC genes reported in the DisGeNET database, 2) DisGeNET_A - 2,096 genes from the DisGeNET database associated with colorectal cancer, intestinal cancer, carcinoma, cancer, neoplasm and 3) Published - 63 CRC genes reported in Peters et al. 2015. Seven CRC gene sets were assessed based on EGA from EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018), *stringentC, flexibleC and inhouseM*. Genes that could not be benchmarked are indicated under name RemainG. For example, 37 CRC-genes identified using EnhancerAtlas EGAs were previously reported in DisGeNET db (Piñero et al. 2017), for CRC. Additional 11 CRC-genes identified using EnhancerAtlas EGAs were reported for intestinal cancer, carcinoma, cancer or neoplasm in DisGeNET db, and 388 genes were not benchmarked. On the other hand, 20 CRC-genes identified using EnhancerAtlas EGAs were reported in Peters et al. 2015.

	DISGENET	DISGENET A	RemainG	Published
EnhancerAtlas	N= 37 [10 %]	N= 11 [3 %]	N= 338 [88 %]	N= 20 [5%]
FlexibleC	N= 6 [12 %]	N= 0 [0 %]	N= 42 [88 %]	N= 3 [6%]
FOCS	N= 3 [19 %]	N= 0 [0 %]	N= 13 [81 %]	N= 2 [12%]
GeneHancer	N= 15 [8 %]	N= 6 [3 %]	N= 179 [90 %]	N= 15 [8%]
InhouseM	N= 11 [6 %]	N= 0 [0 %]	N= 166 [94 %]	N= 3 [2%]
JEME	N= 8 [5 %]	N= 2 [1 %]	N= 136 [93 %]	N= 5 [3%]
StringentC	N= 2 [14 %]	N= 0 [0 %]	N= 12 [86 %]	N= 2 [14%]

Supplementary Table 9. List of genes that were identified by overlapping three CRC benchmark gene sets and CRC-annotated gene sets. Benchmark CRC gene sets correspond to: 1) DisGeNET - 1,676 CRC genes reported in the DisGeNET database, 2) DisGeNET_A - 2,096 genes from the DisGeNET database associated with colorectal cancer, intestinal cancer, carcinoma, cancer, neoplasm and 3) Published - 63 CRC genes reported in Peters et al. 2015. Seven CRC gene sets were assessed based on EGA from EnhancerAtlas (Gao et al. 2016), JEME (Cao et al. 2017), GeneHancer (Fishilevich et al. 2017), FOCS (Hait et al. 2018), *stringentC, flexibleC and inhouseM*.

Method	DISGENET genes	DISGENET ext genes
EA	BMP4, CDKN3, CRLS1, ZMIZ1, MBNL1, ACSL5, CXCL14, DUSP10, MYC, CDH3, CDKN1A, LITAF, SSR1, GALNT12, LNPEP, FOXF1, UTP23, EIF3H, DACT1, DCTN6, DUSP4, IGFBP3, CD9, TNFRSF1A, GAPDH, PTPN6, MLF2, PGC, MDFI, IL6ST, RFC1, DROSHA, CXCR4, ADD1, NEDD9, MIR568, CRAT	CTNNB1, GATA3, MYBL2, CDH1, SRPK1, LAMC1, CDCA3, YEATS4, TLR1, NTMT1, ETV6
Flex	DUSP10, MYC, HOXB9, DVL1, IL6ST, PTGES	
FOCS	HOXB3, HOXB9, DVL1	
GH	BMP4, DUSP10, MYC, FADS1, INCENP, PDCD5, CDKN1A, GALNT12, CD9, PGC, DVL1, LILRB1, PLD1, CXCR4, CRAT	RPL17, CDH1, MORC1, ETV6, YEATS4, NTMT1
IHM	DUSP10, MYC, FOXF1, CDK5RAP3, HOXB9, CORO2A, SDF4, DVL1, IL6ST, DKK4, PTGES	
JEME	GOT1, UTP23, EIF3H, FAM83D, DACT1, CLPTM1L, CRAT, FAS	TERT, CIZ1
Str	DUSP10, DVL1	

Supplementary Table 10. A table of TF that bind or have binding site motifs in the region chr19:33532126-33532564. TFBS indicates transcription factor binding sites that were identified in given region, whereas CHiP-Seq TFBS indicates that peaks for listed TF were identified in a given region

enhancer	TFBS	ChipSeq TFBS
chr19:33532126-33532564	TFAP2A, ELK1, FOXC1, FOXL1, GATA2, GATA3, FOXI1, MZF1_1-4, SOX9, SP1, SPI1, SPIB, SRY, YY1, ETS1, ZNF354C, BRCA1, INSM1, NFIC, SPI1, BATF::JUN, FLI1, FOS, FOSL1, FOSL2, HNF4G, JUN (var.2), JUNB, NRF1, RFX5, SP2, TCF7L2, TFAP2C, EBF1, EGR1, ELK4, FOXA1, HNF4A, NFKB1, SP1, STAT3, THAP1	ARID3A, ATF3, BCL3, BHLHE40, CEBPB, CREB1, ELF1, EP300, ETS1, FOS, FOSL2, FOXA1, FOXA2, GABPA, HDAC2, HNF4A, HNF4G, JUND, MAX, MAZ, MBD4, MXI1, MYBL2, NFIC, NR3C1, RAD21, RCOR1, REST, RFX5, RXRA, SIN3AK20, SIX5, SP1, STAT1, STAT3, TBP, TCF12, TCF7L2, TEAD4, TFAP2A, TFAP2C, USF1, YY1, ZBTB33, ZNF217, ARID3A, ATF3, BCL3, BHLHE40, CEBPB, CREB1, ELF1, EP300, ETS1, FOS, FOSL2, FOXA1, FOXA2, GABPA, HDAC2, HNF4A, HNF4G, JUND, MAX, MAZ, MBD4, MXI1, MYBL2, NFIC, NR3C1, RAD21, RCOR1, REST, RFX5, RXRA, SIN3AK20, SIX5, SP1, STAT1, STAT3, TBP, TCF12, TCF7L2, TEAD4, TFAP2A, TFAP2C, USF1, YY1, ZBTB33, ZNF217

Supplementary Table 11. List of five SNPs that overlapped *stringentC* enhancers, and thus, could be annotated to enhancer-associated LOXL1 gene. Three SNPs (rs62004866, rs150025731, rs28522673) overlapped one enhancer region: chr15:74222688-74224043 (E1); whereas rs4886782 overlapped chr15:74228665-7422857 enhancer (E2) and rs12441130 chr15:74233359-74235313 enhancer (E3). Additional information from the GWAS Catalog was reported as well such as associated phenotype, PMID, etc.

SNP	SNP_Location	SNP_Abb	Enhancer	EnhAbb	GWASGene	predictedGene	PMID	Disease
rs62004866	chr15:74223118	S1	chr15:74222668-74224043	E1	LOXL1	LOXL1	29691431	Hand grip strength
rs150025731	chr15:74223501	S2	chr15:74222668-74224043	E1	LOXL1	LOXL1	27182965	Joint mobility (Beighton score)
rs28522673	chr15:74223716	S3	chr15:74222668-74224043	E1	LOXL1, LOXL1-AS1	LOXL1	29212778	Coronary artery disease
rs4886782	chr15:74228810	S4	chr15:74228665-74228957	E2	LOXL1	LOXL1	25673412	Waist circumference adjusted f or body mass index
rs4886782	chr15:74228810	S4	chr15:74228665-74228957	E2	LOXL1	LOXL1	28448500	Waist circumference adjusted f or body mass index
rs4886782	chr15:74228810	S4	chr15:74228665-74228957	E2	LOXL1	LOXL1	28448500	Waist circumference adjusted f or BMI in active individuals
rs4886782	chr15:74228810	S4	chr15:74228665-74228957	E2	LOXL1	LOXL1	28448500	Waist circumference adjusted f or BMI (joint analysis main ef fects and physical activity in teraction)
rs4886782	chr15:74228810	S4	chr15:74228665-74228957	E2	LOXL1	LOXL1	28443625	Waist circumference adjusted f or BMI (joint analysis main ef fects and smoking interaction)
rs4886782	chr15:74228810	S4	chr15:74228665-74228957	E2	LOXL1	LOXL1	28443625	Waist circumference adjusted f or BMI in non-smokers
rs4886782	chr15:74228810	S4	chr15:74228665-74228957	E2	LOXL1	LOXL1	28443625	Waist circumference adjusted f or BMI (adjusted for smoking b ehaviour)
rs4886782	chr15:74228810	S4	chr15:74228665-74228957	E2	LOXL1	LOXL1	28443625	Waist circumference adjusted f or BMI in smokers
rs12441130	chr15:74234902	S5	chr15:74233359-74235313	E3	NA	LOXL1	30048462	Heel bone mineral density

8.3. Permissions for figures

Figure 1.1.
Figure 1.1. was reproduced with permission (see below) from Springer Nature BV for Article Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.

This is a License Agreement between Inga Patarčić, BIMSB, MDC,Humboldt-Universität ("You") and Springer Nature BV ("Publisher") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier Science & Technology Journals, and the CCC terms and conditions.

All payments must be made in full to CCC.

Order Date 21-Sep-2020
Order license ID 1064478-2
ISSN 0028-0836
Type of Use Republish in a thesis/dissertation
Publisher NATURE PUBLISHING GROUP
Portion Cartoon
LICENSED CONTENT
Publication Title Nature
Article Title Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.
Date 01/01/1869
Language English
Country United Kingdom of Great Britain and Northern Ireland
Rightsholder Springer Nature BV
Publication Type Journal
Start Page 737
End Page 738
Issue 4356
Volume 171
REQUEST DETAILS
Portion Type Cartoon
Number cartoons requested 1
Format (select all that apply) Print, Electronic
Who will republish the content? Academic institution
Duration of Use Life of current edition
Lifetime Unit Quantity Up to 499
Rights Requested Main product
Distribution Worldwide
Translation Original language of publication
Copies for the disabled? No
Minor editing privileges? No
Incidental promotional use? No
Currency EUR
NEW WORK DETAILS
Title Computational mapping of regulatory domains of human genes
Instructor name Inga Patarčić

Institution name BIMSB, MDC
Expected presentation date 2020-12-01
ADDITIONAL DETAILS
Order reference number N/A
The requesting person / organization to appear on the license
Inga Patarčić, BIMSB, MDC, Humboldt-Universität
REUSE CONTENT DETAILS
Title, description or numeric reference of the portion(s) Figure 1
Editor of portion(s) WATSON, J D; CRICK, F H
Volume of serial or monograph 171
Page or page range of portion 737-738
Title of the article/chapter the portion is from Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.
Author of portion(s) WATSON, J D; CRICK, F H
Publication date of portion 1953-04-25

Figure 1.2.

Figure 1.2. was reproduced with permission (see below) from Springer Nature BV for Article From profiles to function in epigenomics.

This is a License Agreement between Inga Patarčić, BIMSB, MDC,Humboldt-Universität ("You") and Springer Nature BV ("Publisher") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier Science & Technology Journals, and the CCC terms and conditions.

All payments must be made in full to CCC.

Order Date 21-Sep-2020
Order license ID 1064478-1
ISSN 1471-0056
Type of Use Republish in a thesis/dissertation
Publisher NATURE PUBLISHING GROUP
Portion Cartoon
LICENSED CONTENT
Publication Title Nature Reviews Genetics
Date 01/01/2000
Language English
Country United Kingdom of Great Britain and Northern Ireland
Rightsholder Springer Nature BV
Publication Type Journal
REQUEST DETAILS
Portion Type Cartoon
Number cartoons requested 1
Format (select all that apply) Print, Electronic
Who will republish the content? Academic institution
Duration of Use Life of current edition

Lifetime Unit Quantity Up to 499
Rights Requested Main product
Distribution Worldwide
Translation Original language of publication
Copies for the disabled? No
Minor editing privileges? No
Incidental promotional use? No
Currency EUR
NEW WORK DETAILS
Title Computational mapping of regulatory domains of human genes
Instructor name Inga Patarčić
Institution name BIMSB, MDC
Expected presentation date 2020-12-01
ADDITIONAL DETAILS
Order reference number N/A
The requesting person / organization to appear on the license
Inga Patarčić, BIMSB, MDC,Humboldt-Universität
REUSE CONTENT DETAILS
Title, description or numeric reference of the portion(s) Box 2 | Profile types and categories
Editor of portion(s) N/A
Volume of serial or monograph 18
Page or page range of portion 55
Title of the article/chapter the portion is from From profiles to function in epigenomics
Author of portion(s) Stefan H. Stricker1,, Anna Köferle and Stephan Beck
Issue, if republishing an article from a serial 1
Publication date of portion 2000-01-01

Figure 1.4.
Figure 1.4. was reproduced with permission (see below) from Elsevier Science & Technology Journals for Article Mapping human epigenomes.

This is a License Agreement between Inga Patarčić, BIMSB, MDC,Humboldt-Universität ("You") and Elsevier Science & Technology Journals ("Publisher") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier Science & Technology Journals, and the CCC terms and conditions.

All payments must be made in full to CCC.

Order Date 13-Aug-2020
Order license ID 1055197-1
ISSN 0092-8674
Type of Use Republish in a thesis/dissertation
Publisher CELL PRESS
Portion Chart/graph/table/figure
LICENSED CONTENT
Publication Title Cell
Article Title Mapping human epigenomes.
Author/Editor National Institute for Medical Research.
Date 01/01/1974

Language English
Country United States of America
Rightsholder Elsevier Science & Technology Journals
Publication Type Journal
Start Page 39
End Page 55
Issue 1
Volume 155
REQUEST DETAILS
Portion Type
Chart/graph/table/figure
Number of charts / graphs / tables / figures requested 1
Format (select all that apply) Print, Electronic
Who will republish the content? Academic institution
Duration of Use Life of current edition
Lifetime Unit Quantity Up to 499
Rights Requested Main product
Distribution Worldwide
Translation Original language of publication
Copies for the disabled? No
Minor editing privileges? No
Incidental promotional use? No
Currency EUR
NEW WORK DETAILS
Title Computational mapping of regulatory domains of human genes
Instructor name Inga Patarčić
Institution name BIMSB, MDC
Expected presentation date 2020-12-01
ADDITIONAL DETAILS
Order reference number N/A
The requesting person / organization to appear on the license
Inga Patarčić, BIMSB, MDC,Humboldt-Universität
REUSE CONTENT DETAILS
Title, description or numeric reference of the portion(s)
Figure 1 Timeline of Sequencing-Based Technologies for Mapping Human Epigenomes
Editor of portion(s) Rivera, Chloe M.; Ren, Bing
Volume of serial or monograph 155
Page or page range of portion 39-55
Title of the article/chapter the portion is from Mapping human epigenomes.
Author of portion(s) Rivera, Chloe M.; Ren, Bing
Issue, if republishing an article from a serial 1
Publication date of portion 2013-09-26

10

Bibliography

1. 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
2. Abbasi, A.A., Paparidis, Z., Malik, S., Goode, D.K., Callaway, H., Elgar, G., and Grzeschik, K.-H. (2007). Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers. PLoS ONE *2*, e366.
3. Adamusiak, T., Burdett, T., Kurbatova, N., Joeri van der Velde, K., Abeygunawardena, N., Antonakaki, D., Kapushesky, M., Parkinson, H., and Swertz, M.A. (2011). OntoCAT--simple ontology search and integration in Java, R and REST/JavaScript. BMC Bioinformatics *12*, 218.
4. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. Molecular biology or the cell. Molecular Biology or the Cell.
5. Albrecht, F., List, M., Bock, C., and Lengauer, T. (2016). DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. Nucleic Acids Res. *44*, W581-6.
6. Allfrey, V.G., Faulkner, R., and Mirsky, A.E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proc Natl Acad Sci USA *51*, 786–794.
7. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
8. Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet. *21*, 71–87.
9. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.
10. Andersson, R., Sandelin, A., and Danko, C.G. (2015). A unified architecture of transcriptional regulatory elements. Trends Genet. *31*, 426–433.
11. Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.
12. Avery, O.T., MacLeod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. *149*, 297–326.
13. Axel, R., Cedar, H., and Felsenfeld, G. (1973). Synthesis of globin ribonucleic acid from duck-reticulocyte chromatin in vitro. Proc Natl Acad Sci USA *70*, 2029–2032.
14. Bae, J.-B. (2013). Perspectives of international human epigenome consortium. Genomics Inform. *11*, 7–14.
15. Bajpai, R., and Nagaraju, G.P. (2017). Specificity protein 1: Its role in colorectal cancer progression and metastasis. Crit. Rev. Oncol. Hematol. *113*, 1–7.
16. Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.
17. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.
18. Bartman, C.R., Hamagami, N., Keller, C.A., Giardine, B., Hardison, R.C., Blobel, G.A., and Raj, A. (2019). Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. Mol. Cell *73*, 519-532.e4.
19. Beadle, G.W., and Tatum, E.L. (1941). Genetic Control of Biochemical Reactions in Neurospora. Proc Natl Acad Sci USA *27*, 499–506.

20. Beagan, J.A., Duong, M.T., Titus, K.R., Zhou, L., Cao, Z., Ma, J., Lachanski, C.V., Gillis, D.R., and Phillips-Cremins, J.E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. *27*, 1139–1152.
21. Bell, O., Tiwari, V.K., Thomä, N.H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. Nat. Rev. Genet. *12*, 554–564.
22. Benayoun, B.A., Pollina, E.A., and Brunet, A. (2015). Epigenetic regulation of ageing: linking environmental inputs to genomic stability. Nat. Rev. Mol. Cell Biol. *16*, 593–610.
23. Bender, M.A., Roach, J.N., Halow, J., Close, J., Alami, R., Bouhassira, E.E., Groudine, M., and Fiering, S.N. (2001). Targeted deletion of 5'HS1 and 5'HS4 of the β-globin locus control region reveals additive activity of the DNaseI hypersensitive sites. Blood *98*, 2022–2027.
24. Berdasco, M., and Esteller, M. (2010). Aberrant epigenetic landscape in cancer: how cellular identity goes awry. Dev. Cell *19*, 698–711.
25. Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA *74*, 3171–3175.
26. Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci USA *99*, 757–762.
27. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol. *28*, 1045–1048.
28. Blackwood, E.M., and Kadonaga, J.T. (1998). Going the distance: a current view of enhancer action. Science *281*, 60–63.
29. Blattler, A., Yao, L., Witt, H., Guo, Y., Nicolet, C.M., Berman, B.P., and Farnham, P.J. (2014). Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. Genome Biol. *15*, 469.
30. Bothma, J.P., Garcia, H.G., Esposito, E., Schlissel, G., Gregor, T., and Levine, M. (2014). Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. Proc Natl Acad Sci USA *111*, 10598–10603.
31. Boveri, T. (1904). Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns. Ergebnisse Über Die Konstitution Der Chromatischen Substanz Des Zellkerns.
32. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. *22*, 1790–1797.
33. Breiman, L. (2001). Random forests. Machine Learning *45*, 5–32.
34. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees (Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software).
35. Brenner, S., Jacob, F., and Meselson, M. (1961). An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. Nature *190*, 576–581.
36. Brunel, H., Gallardo-Chacón, J.-J., Buil, A., Vallverdú, M., Soria, J.M., Caminal, P., and Perera, A. (2010). MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. Bioinformatics *26*, 1811–1818.
37. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.
38. Burch, J.B.E. (2005). Regulation of GATA gene expression during vertebrate development. Seminars in Cell & Developmental Biology *16*, 71–81.
39. Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. Springer Science and Business Media LLC.

40. Calhoun, V.C., Stathopoulos, A., and Levine, M. (2002). Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. Proc Natl Acad Sci USA *99*, 9243–9247.
41. Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.
42. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.
43. Cannavò, E., Khoueiry, P., Garfield, D.A., Geeleher, P., Zichner, T., Gustafson, E.H., Ciglar, L., Korbel, J.O., and Furlong, E.E.M. (2016). Shadow enhancers are pervasive features of developmental regulatory networks. Curr. Biol. *26*, 38–51.
44. Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am. J. Hum. Genet. *86*, 6–22.
45. Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192–197.
46. Cao, Q., Anyansi, C., Hu, X., Xu, L., Xiong, L., Tang, W., Mok, M.T.S., Cheng, C., Fan, X., Gerstein, M., et al. (2017). Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat. Genet. *49*, 1428–1436.
47. Carvajal-Carmona, L.G., Cazier, J.-B., Jones, A.M., Howarth, K., Broderick, P., Pittman, A., Dobbins, S., Tenesa, A., Farrington, S., Prendergast, J., et al. (2011). Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Hum. Mol. Genet. *20*, 2879–2888.
48. Catarino, R.R., and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes Dev. *32*, 202–223.
49. Champoux, J.J. (2001). DNA topoisomerases: structure, function, and mechanism. Annu. Rev. Biochem. *70*, 369–413.
50. Chargaff, E., Lipshitz, R., and Green, C. (1952). Composition of the desoxypentose nucleic acids of four genera of sea-urchin. J. Biol. Chem. *195*, 155–160.
51. Chen, X., and Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics *99*, 323–329.
52. Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018a). Dynamic interplay between enhancer-promoter topology and gene activity. Nat. Genet. *50*, 1296–1303.
53. Chen, H., Li, C., Peng, X., Zhou, Z., Weinstein, J.N., Cancer Genome Atlas Research Network, and Liang, H. (2018b). A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. Cell *173*, 386-399.e12.
54. Chiocchetti, A., Tolosano, E., Hirsch, E., Silengo, L., and Altruda, F. (1997). Green fluorescent protein as a reporter of gene expression in transgenic mice. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression *1352*, 193–202.
55. Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. Cell *12*, 1–8.
56. Cipriano, R., Bryson, B.L., Miskimen, K.L.S., Bartel, C.A., Hernandez-Sanchez, W., Bruntz, R.C., Scott, S.A., Lindsley, C.W., Brown, H.A., and Jackson, M.W. (2014). Hyperactivation of EGFR and downstream effector phospholipase D1 by oncogenic FAM83B. Oncogene *33*, 3298–3306.

57. Clark, S.J., Lee, H.J., Smallwood, S.A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome Biol. *17*, 72.
58. Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A., and Noushmehr, H. (2012). FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. Nucleic Acids Res. *40*, e139.
59. COGENT Study, Houlston, R.S., Webb, E., Broderick, P., Pittman, A.M., Di Bernardo, M.C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., et al. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet. *40*, 1426–1435.
60. Cohen, A.J., Saiakhova, A., Corradin, O., Luppino, J.M., Lovrenert, K., Bartels, C.F., Morrow, J.J., Mack, S.C., Dhillon, G., Beard, L., et al. (2017). Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. Nat. Commun. *8*, 14400.
61. Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.
62. Colbran, L.L., Chen, L., and Capra, J.A. (2017). Short DNA sequence patterns accurately identify broadly active human enhancers. BMC Genomics *18*, 536.
63. Collins, F.S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. Science *300*, 286–290.
64. Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.
65. Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. *46*, 1311–1320.
66. Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. *24*, 1–13.
67. Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D., et al. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. *16*, 123–131.
68. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA *107*, 21931–21936.
69. Crick, F.H. (1958). On protein synthesis. Symp. Soc. Exp. Biol. *12*, 138–163.
70. Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.
71. Crick, F.H., Barnett, L., Brenner, S., and Watts-Tobin, R.J. (1961). General nature of the genetic code for proteins. Nature *192*, 1227–1232.
72. Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. *39*, D691-7.
73. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. *43*, D662-9.
74. Dali, R., and Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. Nucleic Acids Res. *45*, 2994–3005.
75. Dao, L.T.M., Galindo-Albarrán, A.O., Castro-Mondragon, J.A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., et al. (2017).

Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat. Genet. *49*, 1073–1081.
76. Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. J. Mol. Biol. *319*, 1097–1113.
77. Dawson, M.A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. Cell *150*, 12–27.
78. Deblois, G., St-Pierre, J., and Giguère, V. (2013). The PGC-1/ERR signaling axis in cancer. Oncogene *32*, 3483–3490.
79. Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.
80. Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.
81. Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'Shea, C.C., Park, P.J., Ren, B., et al. (2017). The 4D nucleome project. Nature *549*, 219–226.
82. Del Bene, F., Ettwiller, L., Skowronska-Krawczyk, D., Baier, H., Matter, J.-M., Birney, E., and Wittbrodt, J. (2007). In vivo validation of a computationally predicted conserved Ath5 target gene set. PLoS Genet. *3*, 1661–1671.
83. Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G.A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell *149*, 1233–1244.
84. DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet. *14*, 457–460.
85. De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384.
86. Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K.C., Huang, H., Liu, T., Marina, R.J., et al. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat. Methods *14*, 629–635.
87. Diaz-Uriarte, R., and de Andrés, S.A. (2005). Variable selection from random forests: application to gene expression data. ArXiv Preprint Q-Bio/0503025.
88. Dimas, A.S., Stranger, B.E., Beazley, C., Finn, R.D., Ingle, C.E., Forrest, M.S., Ritchie, M.E., Deloukas, P., Tavaré, S., and Dermitzakis, E.T. (2008). Modifier effects between regulatory and protein-coding variation. PLoS Genet. *4*, e1000244.
89. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science *325*, 1246–1250.
90. Ding, J., Gudjonsson, J.E., Liang, L., Stuart, P.E., Li, Y., Chen, W., Weichenthal, M., Ellinghaus, E., Franke, A., Cookson, W., et al. (2010). Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. Am. J. Hum. Genet. *87*, 779–789.
91. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C.C., Taylor, J., Burnett, E., Gut, I., Farrall, M., et al. (2007). A genome-wide association study of global gene expression. Nat. Genet. *39*, 1202–1207.
92. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

93. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.
94. Dorris, D.R., and Struhl, K. (2000). Artificial recruitment of TFIID, but not RNA polymerase II holoenzyme, activates transcription in mammalian cells. Mol. Cell. Biol. *20*, 4350–4358.
95. Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.
96. Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096.
97. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science *314*, 1461–1463.
98. van Duijvenboden, K., de Boer, B.A., Capon, N., Ruijter, J.M., and Christoffels, V.M. (2016). EMERGE: a flexible modelling framework to predict genomic regulatory elements from genomic signatures. Nucleic Acids Res. *44*, e42.
99. Dunipace, L., Ozdemir, A., and Stathopoulos, A. (2011). Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. Development *138*, 4075–4084.
100. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.
101. Ellis, S.E., Gupta, S., Ashar, F.N., Bader, J.S., West, A.B., and Arking, D.E. (2013). RNA-Seq optimization with eQTL gold standards. BMC Genomics *14*, 892.
102. ENCODE Project Consortium (2004). The ENCODE (encyclopedia of DNA elements) project. Science *306*, 636–640.
103. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
104. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.
105. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.
106. Fakhouri, T.H.I., Stevenson, J., Chisholm, A.D., and Mango, S.E. (2010). Dynamic chromatin organization during foregut development mediated by the organ selector gene PHA-4/FoxA. PLoS Genet. *6*.
107. Fearon, E.R. (2011). Molecular genetics of colorectal cancer. Annu. Rev. Pathol. *6*, 479–507.
108. Felsenfeld, G., Boyes, J., Chung, J., Clark, D., and Studitsky, V. (1996). Chromatin structure and gene expression. Proc Natl Acad Sci USA *93*, 9384–9388.
109. Feng, C., Zhang, L., Sun, Y., Li, X., Zhan, L., Lou, Y., Wang, Y., Liu, L., and Zhang, Y. (2018). GDPD5, a target of miR-195-5p, is associated with metastasis and chemoresistance in colorectal cancer. Biomed. Pharmacother. *101*, 945–952.
110. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., et al. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature *260*, 500–507.
111. Fiers, W., Contreras, R., Haegemann, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G., and Ysebaert, M. (1978). Complete nucleotide sequence of SV40 DNA. Nature *273*, 113–120.

112. Finch, J.T., and Klug, A. (1976). Solenoidal model for superstructure in chromatin. Proc Natl Acad Sci USA *73*, 1897–1901.
113. Finn, E.H., Pegoraro, G., Brandão, H.B., Valton, A.-L., Oomen, M.E., Dekker, J., Mirny, L., and Misteli, T. (2019). Extensive heterogeneity and intrinsic variation in spatial genome organization. Cell *176*, 1502-1515.e10.
114. Finver, S.N., Nishikura, K., Finger, L.R., Haluska, F.G., Finan, J., Nowell, P.C., and Croce, C.M. (1988). Sequence analysis of the MYC oncogene involved in the t(8;14)(q24;q11) chromosome translocation in a human leukemia T-cell line indicates that putative regulatory regions are not altered. Proc Natl Acad Sci USA *85*, 3052–3056.
115. Firpi, H.A., Ucar, D., and Tan, K. (2010). Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics *26*, 1579–1586.
116. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) *2017*.
117. Flemming, W. (1880). Beiträge zur kenntniss der zelle und ihrer lebenserscheinungen. Archiv Für Mikroskopische Anatomie *18*, 151–259.
118. Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945-50.
119. Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nat. Methods *14*, 679–685.
120. Frankel, N., Davis, G.K., Vargas, D., Wang, S., Payre, F., and Stern, D.L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature *466*, 490–493.
121. Franklin, R.E., and Gosling, R.G. (1953). Molecular configuration in sodium thymonucleate. Nature *171*, 740–741.
122. Frazer, K.A., Murray, S.S., Schork, N.J., and Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. *10*, 241–251.
123. Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. Nat. Genet. *43*, 513–518.
124. Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R Package Version *1*.
125. Friedman et al. (2001). The Elements of Statistical Learning (New York, NY: Springer New York).
126. Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci USA *89*, 1827–1831.
127. Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science *354*, 769–773.
128. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Patwardhan, T.A., Nguyen, T.H., et al. (2019). Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. BioRxiv.
129. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

130. Gaffney, D.J., Veyrieras, J.-B., Degner, J.F., Pique-Regi, R., Pai, A.A., Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. *13*, R7.
131. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The Post-GWAS Era: From Association to Function. Am. J. Hum. Genet. *102*, 717–730.
132. Gao, T., He, B., Liu, S., Zhu, H., Tan, K., and Qian, J. (2016). EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics *32*, 3543–3551.
133. Garcia-Ramirez, M., Rocchini, C., and Ausio, J. (1995). Modulation of chromatin folding by histone acetylation. J. Biol. Chem. *270*, 17923–17928.
134. Gasperini, M., Findlay, G.M., McKenna, A., Milbank, J.H., Lee, C., Zhang, M.D., Cusanovich, D.A., and Shendure, J. (2017). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. Am. J. Hum. Genet. *101*, 192–205.
135. Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell *176*, 377-390.e19.
136. Gavrilov, A.A., Gushchanskaya, E.S., Strelkova, O., Zhironkina, O., Kireev, I.I., Iarovaia, O.V., and Razin, S.V. (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic Acids Res. *41*, 3563–3575.
137. Gerasimova, A., Chavez, L., Li, B., Seumois, G., Greenbaum, J., Rao, A., Vijayanand, P., and Peters, B. (2013). Predicting cell types and genetic variations contributing to disease by combining GWAS and epigenetic data. PLoS ONE *8*, e54359.
138. Gerster, T., Picard, D., and Schaffner, W. (1986). During B-cell differentiation enhancer activity and transcription rate of immunoglobulin heavy chain genes are high before mRNA accumulation. Cell *45*, 45–52.
139. Gibney, E.R., and Nolan, C.M. (2010). Epigenetics and gene expression. Heredity *105*, 4–13.
140. Gilmour, D.S., Pflugfelder, G., Wang, J.C., and Lis, J.T. (1986). Topoisomerase I interacts with transcribed regions in Drosophila cells. Cell *44*, 401–407.
141. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. *17*, 877–885.
142. Gonzalez-Gay, M.A., Gonzalez-Juanatey, C., and Martin, J. (2005). Rheumatoid arthritis: a disease associated with accelerated atherogenesis. Semin. Arthritis Rheum. *35*, 8–17.
143. Goodson, N. (2002). Coronary artery disease and rheumatoid arthritis. Curr. Opin. Rheumatol. *14*, 115–120.
144. Gorfine, M., Heller, R., and Heller, Y. (2012). Comment on detecting novel associations in large data sets. Science 1–6.
145. Goss, K.H., and Groden, J. (2000). Biology of the adenomatous polyposis coli tumor suppressor. J. Clin. Oncol. *18*, 1967–1979.
146. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.
147. Gregory, T.R. (2011). The evolution of the genome (Elsevier).
148. Grice, E.A., Rochelle, E.S., Green, E.D., Chakravarti, A., and McCallion, A.S. (2005). Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. Hum. Mol. Genet. *14*, 3837–3845.

149. Gross, D.S., and Garrard, W.T. (1988). Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. *57*, 159–197.
150. Groudine, M., Kohwi-Shigematsu, T., Gelinas, R., Stamatoyannopoulos, G., and Papayannopoulou, T. (1983). Human fetal to adult hemoglobin switching: changes in chromatin structure of the beta-globin gene locus. Proc Natl Acad Sci USA *80*, 7551–7555.
151. Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R., Greenside, P., Srivas, R., Phanstiel, D.H., Pekowska, A., et al. (2015). Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell *162*, 1051–1065.
152. GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.
153. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.
154. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.
155. Guo, L., Du, Y., Chang, S., Zhang, K., and Wang, J. (2014a). rSNPBase: a database for curated regulatory SNPs. Nucleic Acids Res. *42*, D1033-9.
156. Guo, X., Zhang, Y., Hu, W., Tan, H., and Wang, X. (2014b). Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation. PLoS ONE *9*, e87446.
157. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
158. Hacia, J.G., Fan, J.B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R.A., Sun, B., Hsie, L., Robbins, C.M., et al. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat. Genet. *22*, 164–167.
159. Hackl, C., Lang, S.A., Moser, C., Mori, A., Fichtner-Feigl, S., Hellerbrand, C., Dietmeier, W., Schlitt, H.J., Geissler, E.K., and Stoeltzing, O. (2010). Activating transcription factor-3 (ATF3) functions as a tumor suppressor in colon cancer and is up-regulated upon heat-shock protein 90 (Hsp90) inhibition. BMC Cancer *10*, 668.
160. Hah, N., Danko, C.G., Core, L., Waterfall, J.J., Siepel, A., Lis, J.T., and Kraus, W.L. (2011). A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell *145*, 622–634.
161. Haines, J.L., Hauser, M.A., Schmidt, S., Scott, W.K., Olson, L.M., Gallins, P., Spencer, K.L., Kwan, S.Y., Noureddine, M., Gilbert, J.R., et al. (2005). Complement factor H variant increases the risk of age-related macular degeneration. Science *308*, 419–421.
162. Hait, T.A., Amar, D., Shamir, R., and Elkon, R. (2018). FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol. *19*, 56.
163. Hamosh, A., Scott, A.F., Amberger, J., Valle, D., and McKusick, V.A. (2000). Online mendelian inheritance in man (OMIM). Hum. Mutat. *15*, 57–61.
164. Hariprakash, J.M., and Ferrari, F. (2019). Computational Biology Solutions to Identify Enhancers-target Gene Pairs. Comput. Struct. Biotechnol. J. *17*, 821–831.

165. Harismendy, O., Notani, D., Song, X., Rahim, N.G., Tanasa, B., Heintzman, N., Ren, B., Fu, X.-D., Topol, E.J., Rosenfeld, M.G., et al. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature *470*, 264–268.
166. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.
167. van der Harst, P., and Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. Circ. Res. *122*, 433–443.
168. Hathaway, N.A., Bell, O., Hodges, C., Miller, E.L., Neel, D.S., and Crabtree, G.R. (2012). Dynamics and memory of heterochromatin in living cells. Cell *149*, 1447–1460.
169. Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L., Kassouf, M.T., Marieke Oudelaar, A.M., Sharpe, J.A., Suciu, M.C., et al. (2016). Genetic dissection of the α-globin super-enhancer in vivo. Nat. Genet. *48*, 895–903.
170. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.
171. Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.
172. Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. *36*, 39–56.
173. He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci USA *111*, E2191-9.
174. He, D., Ma, L., Feng, R., Zhang, L., Jiang, Y., Zhang, Y., and Liu, G. (2015). Analyzing large-scale samples highlights significant association between rs10411210 polymorphism and colorectal cancer. Biomed. Pharmacother. *74*, 164–168.
175. He, H.H., Meyer, C.A., Shin, H., Bailey, S.T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M., et al. (2010). Nucleosome dynamics define transcriptional enhancers. Nat. Genet. *42*, 343–347.
176. He, N., Liu, L., Duan, X., Wang, L., Yuan, D., Jin, T., and Kang, L. (2016). Identification of a shared protective genetic susceptibility locus for colorectal cancer and gastric cancer. Tumour Biol. *37*, 2443–2448.
177. Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat. Biotechnol. *33*, 510–517.
178. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA *106*, 9362–9367.
179. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.
180. Ho, T.K. (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition *1*, 278.
181. Hödl, M., and Basler, K. (2012). Transcription in the absence of histone H3.2 and H3K4 methylation. Curr. Biol. *22*, 2253–2257.
182. Hoerl, A.E., and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics *12*, 55–67.

183. van der Hoeven, F., Zákány, J., and Duboule, D. (1996). Gene transpositions in the HoxD complex reveal a hierarchy of regulatory controls. Cell *85*, 1025–1035.
184. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods *9*, 473–476.
185. Holliday, R. (1990). Mechanisms for the control of gene activity during development. Biol. Rev. Camb. Philos. Soc. *65*, 431–471.
186. Hong, J.-W., Hendrix, D.A., and Levine, M.S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.
187. Hotchkiss, R.D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J. Biol. Chem. *175*, 315–332.
188. Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262–1278.
189. Ibn-Salem, J., Köhler, S., Love, M.I., Chung, H.-R., Huang, N., Hurles, M.E., Haendel, M., Washington, N.L., Smedley, D., Mungall, C.J., et al. (2014). Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol. *15*, 423.
190. Innocenti, F., Cooper, G.M., Stanaway, I.B., Gamazon, E.R., Smith, J.D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y.S., Moloney, C., et al. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet. *7*, e1002078.
191. Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., and Lauer, M.S. (2010). High-Dimensional Variable Selection for Survival Data. J. Am. Stat. Assoc. *105*, 205–217.
192. Iyer, L.M., Zhang, D., and Aravind, L. (2016). Adenine methylation in eukaryotes: Apprehending the complex evolutionary history and functional potential of an epigenetic modification. Bioessays *38*, 27–40.
193. Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.
194. Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. Cell *167*, 1883-1896.e15.
195. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369-1384.e19.
196. Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074–1080.
197. Jiao, S., Peters, U., Berndt, S., Brenner, H., Butterbach, K., Caan, B.J., Carlson, C.S., Chan, A.T., Chang-Claude, J., Chanock, S., et al. (2014). Estimating the heritability of colorectal cancer. Hum. Mol. Genet. *23*, 3898–3905.
198. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.
199. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.
200. Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.
201. Joshi, A. (2014). Mammalian transcriptional hotspots are enriched for tissue specific enhancers near cell type specific highly expressed genes and are predicted to act as transcriptional activator hubs. BMC Bioinformatics *15*, 412.

202. Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.
203. Kanehisa, M. (2002). The KEGG database. Novartis Found. Symp. *247*, 91–101; discussion 101.
204. Kang, B.W., Jeon, H.-S., Chae, Y.S., Lee, S.J., Park, J.Y., Choi, J.E., Park, J.S., Choi, G.S., and Kim, J.G. (2015). Association between GWAS-identified genetic variations and disease prognosis for patients with colorectal cancer. PLoS ONE *10*, e0119649.
205. Kantorovitz, M.R., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G.E., Göttgens, B., Halfon, M.S., and Sinha, S. (2009). Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. Dev. Cell *17*, 568–579.
206. Kaplan, T., Li, X.-Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D., and Eisen, M.B. (2011). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet. *7*, e1001290.
207. Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484–1488.
208. Kearns, N.A., Pham, H., Tabak, B., Genga, R.M., Silverstein, N.J., Garber, M., and Maehr, R. (2015). Functional annotation of native enhancers with a Cas9-histone demethylase fusion. Nat. Methods *12*, 401–403.
209. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA *111*, 6131–6138.
210. Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res. *17*, 1919–1931.
211. Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811.
212. Kho, D.H., Bae, J.A., Lee, J.H., Cho, H.J., Cho, S.H., Lee, J.H., Seo, Y.W., Ahn, K.Y., Chung, I.J., and Kim, K.K. (2009). KITENIN recruits Dishevelled/PKC delta to form a functional complex and controls the migration and invasiveness of colorectal cancer cells. Gut *58*, 509–519.
213. Kim, S.K. (2018). Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. PLoS ONE *13*, e0200785.
214. Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.
215. Kleftogiannis, D., Kalnis, P., and Bajic, V.B. (2015). DEEP: a general computational framework for predicting enhancers. Nucleic Acids Res. *43*, e6.
216. Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science *184*, 868–871.
217. Kossel, A. (1911). The chemical composition of the cell. The Chemical Composition of the Cell.
218. Kothary, R., Clapoff, S., Darling, S., Perry, M.D., Moran, L.A., and Rossant, J. (1989). Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. Development *105*, 707–714.

219. Kulaeva, O.I., Nizovtseva, E.V., Polikanov, Y.S., Ulianov, S.V., and Studitsky, V.M. (2012). Distant activation of transcription: mechanisms of enhancer action. Mol. Cell. Biol. *32*, 4892–4897.
220. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90-7.
221. Kvon, E.Z. (2015). Using transgenic reporter assays to functionally characterize enhancers in animals. Genomics *106*, 185–192.
222. Kvon, E.Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J.O., Pagani, M., Schernhuber, K., Dickson, B.J., and Stark, A. (2014). Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature *512*, 91–95.
223. Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci USA *109*, 19498–19503.
224. Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.
225. Lajoie, B.R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods *72*, 65–75.
226. Landry, J.-R., Bonadies, N., Kinston, S., Knezevic, K., Wilson, N.K., Oram, S.H., Janes, M., Piltz, S., Hammett, M., Carter, J., et al. (2009). Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors. Blood *113*, 5783–5792.
227. Larson, N.B., McDonnell, S., French, A.J., Fogarty, Z., Cheville, J., Middha, S., Riska, S., Baheti, S., Nair, A.A., Wang, L., et al. (2015). Comprehensively evaluating cis-regulatory variation in the human prostate transcriptome by using gene-level allele-specific expression. Am. J. Hum. Genet. *96*, 869–882.
228. Lathrop, G.M., Lalouel, J.M., Julier, C., and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA *81*, 3443–3446.
229. LeBleu, V.S., O'Connell, J.T., Gonzalez Herrera, K.N., Wikman, H., Pantel, K., Haigis, M.C., de Carvalho, F.M., Damascena, A., Domingos Chinen, L.T., Rocha, R.M., et al. (2014). PGC-1α mediates mitochondrial biogenesis and oxidative phosphorylation in cancer cells to promote metastasis. Nat. Cell Biol. *16*, 992–1003, 1.
230. Leder, P., and Nirenberg, M.W. (1964). Rna codewords and protein synthesis, iii. on the nucleotide sequence of a cysteine and a leucine rna codeword. Proceedings of the National Academy of Sciences *52*, 1521–1529.
231. Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.
232. Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet. *47*, 955–961.
233. Lemon, B., and Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. *14*, 2551–2569.
234. Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat. Rev. Genet. *13*, 233–245.
235. Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

236. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y., et al. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature *518*, 350–354.
237. Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147–151.
238. Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping back to leap forward: transcription enters a new era. Cell *157*, 13–25.
239. Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R News *2*, 18–22.
240. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.
241. Lim, U., Wilkens, L.R., Monroe, K.R., Caberto, C., Tiirikainen, M., Cheng, I., Park, S.L., Stram, D.O., Henderson, B.E., Kolonel, L.N., et al. (2012). Susceptibility variants for obesity and colorectal cancer risk: the multiethnic cohort and PAGE studies. Int. J. Cancer *131*, E1038-43.
242. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.
243. Liu, C.-Y., Wu, M.C., Chen, F., Ter-Minassian, M., Asomaning, K., Zhai, R., Wang, Z., Su, L., Heist, R.S., Kulke, M.H., et al. (2010). A Large-scale genetic association study of esophageal adenocarcinoma risk. Carcinogenesis *31*, 1259–1263.
244. Liu, X., Zhao, Y., Gao, J., Pawlyk, B., Starcher, B., Spencer, J.A., Yanagisawa, H., Zuo, J., and Li, T. (2004). Elastic fiber homeostasis requires lysyl oxidase-like 1 protein. Nat. Genet. *36*, 178–182.
245. Li, M.J., Wang, L.Y., Xia, Z., Sham, P.C., and Wang, J. (2013). GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucleic Acids Res. *41*, W150-8.
246. Li, Y., Shi, W., and Wasserman, W.W. (2018). Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics *19*, 202.
247. Lü, B., Fang, Y., Xu, J., Wang, L., Xu, F., Xu, E., Huang, Q., and Lai, M. (2008). Analysis of SOX9 expression in colorectal cancer. Am. J. Clin. Pathol. *130*, 897–904.
248. Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251–260.
249. Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.
250. MacNair, L., Xiao, S., Miletic, D., Ghani, M., Julien, J.-P., Keith, J., Zinman, L., Rogaeva, E., and Robertson, J. (2016). MTHFSD and DDX58 are novel RNA-binding proteins abnormally regulated in amyotrophic lateral sclerosis. Brain *139*, 86–100.
251. Madrigal, P., and Krajewski, P. (2012). Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. Front. Genet. *3*, 230.
252. Mairhofer, M., Steiner, M., Salzer, U., and Prohaska, R. (2009). Stomatin-like protein-1 interacts with stomatin and is targeted to late endosomes. J. Biol. Chem. *284*, 29218–29229.
253. Malan, T.P., and McClure, W.R. (1984). Dual promoter control of the Escherichia coli lactose operon. Cell *39*, 173–180.

254. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.
255. Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. *118*, 1590–1605.
256. Manosas, M., Xi, X.G., Bensimon, D., and Croquette, V. (2010). Active and passive mechanisms of helicases. Nucleic Acids Res. *38*, 5518–5526.
257. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751–753.
258. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.
259. Matana, A., Brdar, D., Torlak, V., Boutin, T., Popović, M., Gunjača, I., Kolčić, I., Boraska Perica, V., Punda, A., Polašek, O., et al. (2018). Genome-wide meta-analysis identifies novel loci associated with parathyroid hormone level. Mol. Med. *24*, 15.
260. Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C., Chou, A., Ienasescu, H., et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. *42*, D142-7.
261. Mathelier, A., Shi, W., and Wasserman, W.W. (2015). Identification of altered cis-regulatory elements in human disease. Trends Genet. *31*, 67–76.
262. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.
263. Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc Natl Acad Sci USA *74*, 560–564.
264. McInnes, I.B., and Schett, G. (2011). The pathogenesis of rheumatoid arthritis. N. Engl. J. Med. *365*, 2205–2219.
265. McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., et al. (2001). A physical map of the human genome. Nature *409*, 934–941.
266. Mendel, G. (1866). Versuche über Pflanzen-Hybriden / (Brünn : Im Verlage des Vereines,).
267. Merli, C., Bergstrom, D.E., Cygan, J.A., and Blackman, R.K. (1996). Promoter specificity mediates the independent regulation of neighboring genes. Genes Dev. *10*, 1260–1270.
268. Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.
269. Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R., and Furlong, E.E.M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev. *32*, 42–57.
270. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.
271. Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature *237*, 82–88.
272. Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. *8*, 1551–1566.

273. Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques *45*, 81–94.
274. Mori, K., Toiyama, Y., Otake, K., Fujikawa, H., Saigusa, S., Hiro, J., Kobayashi, M., Ohi, M., Tanaka, K., Inoue, Y., et al. (2017). Proteomics analysis of differential protein expression identifies heat shock protein 47 as a predictive marker for lymph node metastasis in patients with colorectal cancer. Int. J. Cancer *140*, 1425–1435.
275. Mount, D.W., and Mount, D.W. (2001). Bioinformatics: sequence and genome analysis (New York:: Cold spring harbor laboratory press).
276. Müeller-Storm, H.P., Sogo, J.M., and Schaffner, W. (1989). An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. Cell *58*, 767–777.
277. Murakawa, Y., Yoshihara, M., Kawaji, H., Nishikawa, M., Zayed, H., Suzuki, H., Fantom Consortium, and Hayashizaki, Y. (2016). Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. Trends Genet. *32*, 76–88.
278. Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat. Genet. *39*, 1507–1511.
279. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.
280. Nagasawa, A., Kudoh, J., Noda, S., Mashima, Y., Wright, A., Oguchi, Y., and Shimizu, N. (1999). Human and mouse ISLR (immunoglobulin superfamily containing leucine-rich repeat) genes: genomic structure and tissue expression. Genomics *61*, 37–43.
281. Nagaraju. G.P., Bramhachari (2017). Role of transcription factors in gastrointestinal malignancies (Singapore: Springer Singapore).
282. Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science *347*, 1017–1021.
283. Narlikar, L., Sakabe, N.J., Blanski, A.A., Arimura, F.E., Westlund, J.M., Nobrega, M.A., and Ovcharenko, I. (2010). Genome-wide discovery of human heart enhancers. Genome Res. *20*, 381–392.
284. Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the mitotic chromosome. Science *342*, 948–953.
285. Niittymäki, I., Tuupanen, S., Li, Y., Järvinen, H., Mecklin, J.-P., Tomlinson, I.P.M., Houlston, R.S., Karhu, A., and Aaltonen, L.A. (2011). Systematic search for enhancer elements and somatic allelic imbalance at seven low-penetrance colorectal cancer predisposition loci. BMC Med. Genet. *12*, 23.
286. Nirenberg, M.W., and Matthaei, J.H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci USA *47*, 1588–1602.
287. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.
288. Nyhan, W.L., and Sakati, N.A. (1987). Diagnostic recognition of genetic disease.
289. O'Connor, T., Bodén, M., and Bailey, T.L. (2017). CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data. Nucleic Acids Res. *45*, e19.
290. O'Kane, C.J., and Gehring, W.J. (1987). Detection in situ of genomic regulatory elements in Drosophila. Proc Natl Acad Sci USA *84*, 9123–9127.

291. O'Sullivan, J.M., Hendy, M.D., Pichugina, T., Wake, G.C., and Langowski, J. (2013). The statistical-mechanics of chromosome conformation capture. Nucleus *4*, 390–398.
292. Ogryzko, V.V., Schiltz, R.L., Russanova, V., Howard, B.H., and Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. Cell *87*, 953–959.
293. Ohtsuki, S., Levine, M., and Cai, H.N. (1998). Different core promoters possess distinct regulatory activities in the Drosophila embryo. Genes Dev. *12*, 547–556.
294. Olins, A.L., and Olins, D.E. (1974). Spheroid chromatin units (v bodies). Science *183*, 330–332.
295. Ono, H., Iizumi, Y., Goi, W., Sowa, Y., Taguchi, T., and Sakai, T. (2017). Ribosomal protein S3 regulates XIAP expression independently of the NF-κB pathway in breast cancer cells. Oncol. Rep. *38*, 3205–3210.
296. Panousis, N.I., Gutierrez-Arcelus, M., Dermitzakis, E.T., and Lappalainen, T. (2014). Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. Genome Biol. *15*, 467.
297. Papanikolaou, N., Pavlopoulos, G.A., Theodosiou, T., and Iliopoulos, I. (2015). Protein-protein interaction predictions using text mining methods. Methods *74*, 47–53.
298. Pardee, A.B., Jacob, F., and Monod, J. (1959). The genetic control and cytoplasmic expression of "Inducibility" in the synthesis of β-galactosidase by E. coli. J. Mol. Biol. *1*, 165–178.
299. Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.
300. Park, M.Y., and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. Biostatistics *9*, 30–50.
301. Pasutto, F., Zenkel, M., Hoja, U., Berner, D., Uebe, S., Ferrazzi, F., Schödel, J., Liravi, P., Ozaki, M., Paoli, D., et al. (2017). Pseudoexfoliation syndrome-associated genetic variants affect transcription factor binding and alternative splicing of LOXL1. Nat. Commun. *8*, 15466.
302. Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. *27*, 1173–1175.
303. Pazin, M.J., and Kadonaga, J.T. (1997). SWI2/SNF2 and related proteins: ATP-driven motors that disrupt protein-DNA interactions? Cell *88*, 737–740.
304. Pearson, K. (1895). Notes on Regression and Inheritance in the Case of Two Parents Proceedings of the Royal Society of London, 58, 240-242.
305. Pengelly, A.R., Copur, Ö., Jäckle, H., Herzig, A., and Müller, J. (2013). A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. Science *339*, 698–699.
306. Peters, U., Bien, S., and Zubair, N. (2015). Genetic architecture of colorectal cancer. Gut *64*, 1623–1636.
307. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
308. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.
309. Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. *45*, D833–D839.

310. Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. *21*, 447–455.
311. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., and Jensen, L.J. (2015). DISEASES: text mining and data integration of disease-gene associations. Methods *74*, 83–89.
312. Png, C.W., Weerasooriya, M., Guo, J., James, S.J., Poh, H.M., Osato, M., Flavell, R.A., Dong, C., Yang, H., and Zhang, Y. (2016). DUSP10 regulates intestinal epithelial cell growth and colorectal tumorigenesis. Oncogene *35*, 206–217.
313. Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. *16*, 245–257.
314. Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M., et al. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat. Genet. *41*, 882–884.
315. Ptashne, M., Backman, K., Humayun, M.Z., Jeffrey, A., Maurer, R., Meyer, B., and Sauer, R.T. (1976). Autoregulation and function of a repressor in bacteriophage lambda. Science *194*, 156–161.
316. Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins *63*, 490–500.
317. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.
318. Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., and Ren, B. (2013). RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput. Biol. *9*, e1002968.
319. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.
320. Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015). High-throughput sequencing technologies. Mol. Cell *58*, 586–597.
321. Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J.M., Kim, J., Lawrence, M.S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M., et al. (2017). Recurrent and functional regulatory mutations in breast cancer. Nature *547*, 55–60.
322. Riethoven, J.-J.M. (2010). Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. Methods Mol. Biol. *674*, 33–42.
323. Rivenbark, A.G., Stolzenburg, S., Beltran, A.S., Yuan, X., Rots, M.G., Strahl, B.D., and Blancafort, P. (2012). Epigenetic reprogramming of cancer cells via targeted DNA methylation. Epigenetics *7*, 350–360.
324. Rivera, C.M., and Ren, B. (2013). Mapping human epigenomes. Cell *155*, 39–55.
325. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
326. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods *4*, 651–657.
327. Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nat. Rev. Genet. *7*, 862–872.

328. Roeder, R.G. (1996). The role of general initiation factors in transcription by RNA polymerase II. Trends Biochem. Sci. *21*, 327–335.
329. Roeder, R.G., and Rutter, W.J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224*, 234–237.
330. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. *242*, 84–89.
331. Ronald, J., Brem, R.B., Whittle, J., and Kruglyak, L. (2005). Local regulatory variation in Saccharomyces cerevisiae. PLoS Genet. *1*, e25.
332. Ronen, J., Hayat, S., and Akalin, A. (2019). Evaluation of colorectal cancer subtypes and cell lines using deep learning. Life Sci. Alliance *2*.
333. Roy, S., Siahpirani, A.F., Chasman, D., Knaack, S., Ay, F., Stewart, R., Wilson, M., and Sridharan, R. (2015). A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res. *43*, 8694–8712.
334. Sabarís, G., Laiker, I., Preger-Ben Noon, E., and Frankel, N. (2019). Actors with Multiple Roles: Pleiotropic Enhancers and the Paradigm of Enhancer Modularity. Trends Genet. *35*, 423–433.
335. Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA *74*, 5463–5467.
336. Schaffner, W. (2015). Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. Biol. Chem. *396*, 311–327.
337. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.
338. Schleif, R. (1992). DNA looping. Annu. Rev. Biochem. *61*, 199–223.
339. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et al. (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. *17*, 2042–2059.
340. Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. *25*, 582–597.
341. Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.
342. Schork, A.J., Won, H., Appadurai, V., Nudel, R., Gandal, M., Delaneau, O., Revsbech Christiansen, M., Hougaard, D.M., Bækved-Hansen, M., Bybjerg-Grauholm, J., et al. (2019). A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. Nat. Neurosci. *22*, 353–361.
343. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. *19*, 212–219.
344. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.
345. Schwartz, R.J., and Olson, E.N. (1999). Building the heart piece by piece: modularity of cis-elements regulating Nkx2-5 transcription. Development *126*, 4187–4192.
346. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

347. Sharpe, J., Nonchev, S., Gould, A., Whiting, J., and Krumlauf, R. (1998). Selectivity, sharing and competitive interactions in the regulation of Hoxb genes. EMBO J. *17*, 1788–1798.
348. Sheffield, N.C., Thurman, R.E., Song, L., Safi, A., Stamatoyannopoulos, J.A., Lenhard, B., Crawford, G.E., and Furey, T.S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. *23*, 777–788.
349. Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.
350. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., et al. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature *555*, 611–616.
351. Shrine, N., Guyatt, A.L., Erzurumluoglu, A.M., Jackson, V.E., Hobbs, B.D., Melbourne, C.A., Batini, C., Fawcett, K.A., Song, K., Sakornsakolpat, P., et al. (2019). New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat. Genet. *51*, 481–493.
352. Siddique, A.N., Nunna, S., Rajavelu, A., Zhang, Y., Jurkowska, R.Z., Reinhardt, R., Rots, M.G., Ragozin, S., Jurkowski, T.P., and Jeltsch, A. (2013). Targeted methylation and gene silencing of VEGF-A in human cells by using a designed Dnmt3a-Dnmt3L single-chain fusion protein with increased DNA methylation activity. J. Mol. Biol. *425*, 479–491.
353. Sillén, A., Andrade, J., Lilius, L., Forsell, C., Axelman, K., Odeberg, J., Winblad, B., and Graff, C. (2008). Expanded high-resolution genetic study of 109 Swedish families with Alzheimer's disease. Eur. J. Hum. Genet. *16*, 202–208.
354. Silva, T.C., Coetzee, S.G., Gull, N., Yao, L., Hazelett, D.J., Noushmehr, H., Lin, D.-C., and Berman, B.P. (2019). ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. Bioinformatics *35*, 1974–1977.
355. Simonet, W.S., Bucay, N., Pitas, R.E., Lauer, S.J., and Taylor, J.M. (1991). Multiple tissue-specific elements control the apolipoprotein E/C-I gene locus in transgenic mice. J. Biol. Chem. *266*, 8651–8654.
356. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. *38*, 1348–1354.
357. Singh, S., Yang, Y., Póczos, B., and Ma, J. (2019). Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant. Biol. *7*, 122–137.
358. Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annu. Rev. Biochem. *72*, 449–479.
359. Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). BioMart--biological queries made easy. BMC Genomics *10*, 22.
360. Smemo, S., Tena, J.J., Kim, K.-H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371–375.
361. Smith, E., and Shilatifard, A. (2014). Enhancer biology and enhanceropathies. Nat. Struct. Mol. Biol. *21*, 210–219.
362. Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J. *32*, 3119–3129.

363. Solomon, M.J., Larsen, P.L., and Varshavsky, A. (1988). Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. Cell *53*, 937–947.
364. Spielmann, M., and Mundlos, S. (2013). Structural variations, the regulatory landscape of the genome and their alteration in human disease. Bioessays *35*, 533–543.
365. Spielmann, M., Brancati, F., Krawitz, P.M., Robinson, P.N., Ibrahim, D.M., Franke, M., Hecht, J., Lohan, S., Dathe, K., Nardone, A.M., et al. (2012). Homeotic arm-to-leg transformation associated with genomic rearrangements at the PITX1 locus. Am. J. Hum. Genet. *91*, 629–635.
366. Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., et al. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. Nat. Genet. *50*, 238–249.
367. Stern, M., Jensen, R., and Herskowitz, I. (1984). Five SWI genes are required for expression of the HO gene in yeast. J. Mol. Biol. *178*, 853–868.
368. Stricker, S.H., Köferle, A., and Beck, S. (2017). From profiles to function in epigenomics. Nat. Rev. Genet. *18*, 51–66.
369. Stunnenberg, H.G., International Human Epigenome Consortium, and Hirst, M. (2016). The international human epigenome consortium: A blueprint for scientific collaboration and discovery. Cell *167*, 1145–1149.
370. Styrkarsdottir, U., Lund, S.H., Thorleifsson, G., Zink, F., Stefansson, O.A., Sigurdsson, J.K., Juliusson, K., Bjarnadottir, K., Sigurbjornsdottir, S., Jonsson, S., et al. (2018). Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis. Nat. Genet. *50*, 1681–1687.
371. Sur, I., and Taipale, J. (2016). The role of enhancers in cancer. Nat. Rev. Cancer *16*, 483–493.
372. Sutton, W.S. (1902). On the morphology of the chromosome group in Brachystola magna. Reson. *14*, 398–411.
373. Sutton, W.S. (1903). The chromosomes in heredity. Biol. Bull. *4*, 231–250.
374. Su, W., Jackson, S., Tjian, R., and Echols, H. (1991). DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. Genes Dev. *5*, 820–826.
375. Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. Genome Res. *24*, 390–400.
376. Székely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. Ann. Statist. *35*, 2769–2794.
377. Szutorisz, H., Dillon, N., and Tora, L. (2005). The role of enhancers as centres for general transcription factor recruitment. Trends Biochem. Sci. *30*, 593–599.
378. Tak, Y.G., and Farnham, P.J. (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin *8*, 57.
379. Tang, H., Sun, X., Reinberg, D., and Ebright, R.H. (1996). Protein-protein interactions in eukaryotic transcription initiation: structure of the preinitiation complex. Proc Natl Acad Sci USA *93*, 1119–1124.
380. Tang, W., Zhou, W., Xiang, L., Wu, X., Zhang, P., Wang, J., Liu, G., Zhang, W., Peng, Y., Huang, X., et al. (2019). The p300/YY1/miR-500a-5p/HDAC2 signalling axis regulates cell proliferation in human colorectal cancer. Nat. Commun. *10*, 663.
381. Tan, G., and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics *32*, 1555–1556.

382. Taub, F.E., DeLeo, J.M., and Thompson, E.B. (1983). Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs. DNA *2*, 309–327.
383. Taunton, J., Hassig, C.A., and Schreiber, S.L. (1996). A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. Science *272*, 408–411.
384. Tenesa, A., and Dunlop, M.G. (2009). New insights into the aetiology of colorectal cancer from genome-wide association studies. Nat. Rev. Genet. *10*, 353–358.
385. Terasaki, H., Saitoh, T., Shiokawa, K., and Katoh, M. (2002). Frizzled-10, up-regulated in primary colorectal cancer, is a positive regulator of the WNT - β-catenin - TCF signaling pathway. Int. J. Mol. Med.
386. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinformatics *14*, 178–192.
387. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.
388. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) *58*, 267–288.
389. Tikhonov, A.N., Goncharsky, A.V., Stepanov, V.V., and Yagola, A.G. (1995). Numerical Methods for the Solution of Ill-Posed Problems (Dordrecht: Springer Netherlands).
390. Tikkanen, E., Gustafsson, S., Amar, D., Shcherbina, A., Waggott, D., Ashley, E.A., and Ingelsson, E. (2018). Biological insights into muscular strength: genetic findings in the UK biobank. Sci. Rep. *8*, 6451.
391. Timofeeva, M.N., Kinnersley, B., Farrington, S.M., Whiffin, N., Palles, C., Svinti, V., Lloyd, A., Gorman, M., Ooi, L.-Y., Hosking, F., et al. (2015). Recurrent coding sequence variation explains only A small fraction of the genetic architecture of colorectal cancer. Sci. Rep. *5*, 16286.
392. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol. Cell *10*, 1453–1465.
393. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.
394. Tsujimura, T., Klein, F.A., Langenfeld, K., Glaser, J., Huber, W., and Spitz, F. (2015). A discrete transition zone organizes the topological and regulatory autonomy of the adjacent tfap2c and bmp7 genes. PLoS Genet. *11*, e1004897.
395. Tsymbal, A., Pechenizkiy, M., and Cunningham, P. (2005). Diversity in search strategies for ensemble feature selection. Information Fusion *6*, 83–98.
396. Varley, K.E., Gertz, J., Bowling, K.M., Parker, S.L., Reddy, T.E., Pauli-Behn, F., Cross, M.K., Williams, B.A., Stamatoyannopoulos, J.A., Crawford, G.E., et al. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. *23*, 555–567.
397. Veltman, J.A., and Brunner, H.G. (2012). De novo mutations in human genetic disease. Nat. Rev. Genet. *13*, 565–575.
398. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.
399. Verdin, E., and Ott, M. (2015). 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. Nat. Rev. Mol. Cell Biol. *16*, 258–264.

400. Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G., and Higgs, D.R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. EMBO J. *26*, 2041–2051.
401. Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. *10*, 1297–1309.
402. Vijay, N., Poelstra, J.W., Künstner, A., and Wolf, J.B.W. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol. Ecol. *22*, 620–634.
403. Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.-M., et al. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature *439*, 871–874.
404. Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. *35*, D88-92.
405. Visel, A., Akiyama, J.A., Shoukry, M., Afzal, V., Rubin, E.M., and Pennacchio, L.A. (2009). Functional autonomy of distant-acting human enhancers. Genomics *93*, 509–513.
406. Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.
407. Waddington, C.H. (1959). Canalization of development and genetic assimilation of acquired characters. Nature *183*, 1654–1655.
408. Wagner, E.K., Hewlett, M.J., Bloom, D.C., and Camerini, D. (1999). Basic virology (Malden MA: Blackwell Science).
409. Wakabayashi, A., Ulirsch, J.C., Ludwig, L.S., Fiorini, C., Yasuda, M., Choudhuri, A., McDonel, P., Zon, L.I., and Sankaran, V.G. (2016). Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. Proc Natl Acad Sci USA *113*, 4434–4439.
410. Waldeyer, W. (1888). Ueber Karyokinese und ihre Beziehungen zu den Befruchtungsvorgängen. Archiv Für Mikroskopische Anatomie *32*, 1–122.
411. Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., and Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. Front. Genet. *4*, 270.
412. Walsh, C. (2006) Posttranslational modification of proteins : expanding... - Google Scholar. Posttranslational Modification of Proteins: Expanding Nature's Inventory.
413. Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A., et al. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature *474*, 390–394.
414. Wang, J., Dai, X., Berry, L.D., Cogan, J.D., Liu, Q., and Shyr, Y. (2019). HACER: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Res. *47*, D106–D112.
415. Wang, Q., Peng, R., Wang, B., Wang, J., Yu, W., Liu, Y., and Shi, G. (2018). Transcription factor KLF13 inhibits AKT activation and suppresses the growth of prostate carcinoma cells. Cancer Biomark. *22*, 533–541.
416. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40*, D930-4.
417. Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.
418. Weber, F., and Schaffner, W. (1985). Simian virus 40 enhancer increases RNA polymerase density within the linked gene. Nature *315*, 75–77.

419. Weintraub, H., and Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. Science *193*, 848–856.
420. Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell *171*, 1573-1588.e28.
421. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001-6.
422. Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. *45*, 1238–1243.
423. Whalen, S., Truty, R.M., and Pollard, K.S. (2015). Protein binding and methylation on looping chromatin accurately predict distal regulatory interactions. BioRxiv.
424. Whalen, S., Truty, R.M., and Pollard, K.S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet. *48*, 488–496.
425. White, R.J., Collins, J.E., Sealy, I.M., Wali, N., Dooley, C.M., Digby, Z., Stemple, D.L., Murphy, D.N., Billis, K., Hourlier, T., et al. (2017). A high-resolution mRNA expression time course of embryonic development in zebrafish. Elife *6*.
426. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.
427. Wiench, M., John, S., Baek, S., Johnson, T.A., Sung, M.-H., Escobar, T., Simmons, C.A., Pearce, K.H., Biddie, S.C., Sabo, P.J., et al. (2011). DNA methylation status predicts cell type-specific enhancer activity. EMBO J. *30*, 3028–3039.
428. Winkler, H. (1920). Verbreitung und ursache der parthenogenesis im pflanzen- und tierreiche.
429. Winter, G.E., Mayer, A., Buckley, D.L., Erb, M.A., Roderick, J.E., Vittori, S., Reyes, J.M., di Iulio, J., Souza, A., Ott, C.J., et al. (2017). BET bromodomain proteins function as master transcription elongation factors independent of CDK9 recruitment. Mol. Cell *67*, 5-18.e19.
430. de Wit, E., Bouwman, B.A.M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J.A.M., Krijger, P.H.L., Festuccia, N., Nora, E.P., Welling, M., et al. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227–231.
431. Woodcock, C.L., Safer, J.P., and Stanchfield, J.E. (1976). Structural repeating units in chromatin. I. Evidence for their general occurrence. Exp. Cell Res. *97*, 101–110.
432. Workman, J.L., and Kingston, R.E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. Annu. Rev. Biochem. *67*, 545–579.
433. Wreczycka, K., Gosdschan, A., Yusuf, D., Grüning, B., Assenov, Y., and Akalin, A. (2017). Strategies for analyzing bisulfite sequencing data. J. Biotechnol. *261*, 105–115.
434. Wreczycka, K., Franke, V., Uyar, B., Wurmus, R., Bulut, S., Tursun, B., and Akalin, A. (2019). HOT or not: examining the basis of high-occupancy target regions. Nucleic Acids Res. *47*, 5735–5745.
435. Wu, R., Yun, Q., Zhang, J., and Bao, J. (2019). Downregulation of KLF13 through DNMT1-mediated hypermethylation promotes glioma cell proliferation and invasion. Onco Targets Ther *12*, 1509–1520.
436. Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics *25*, 714–721.

437. Wu, Y., Qi, T., Wang, H., Zhang, F., Zheng, Z., Phillips-Cremins, J.E., Deary, I.J., McRae, A.F., Wray, N.R., Zeng, J., et al. (2020). Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data. Nat. Commun. *11*, 2061.
438. Xie, X., Ma, W., Songyang, Z., Luo, Z., Huang, J., Dai, Z., and Xiong, Y. (2016). CCSI: a database providing chromatin-chromatin spatial interaction information. Database (Oxford) *2016*.
439. Xing, Y., Xu, Y., Chen, Y., Jeffrey, P.D., Chao, Y., Lin, Z., Li, Z., Strack, S., Stock, J.B., and Shi, Y. (2006). Structure of protein phosphatase 2A core enzyme bound to tumor-inducing toxins. Cell *127*, 341–353.
440. Yang, P., Hwa Yang, Y., B. Zhou, B., and Y. Zomaya, A. (2010). A Review of Ensemble Methods in Bioinformatics. Curr. Bioinform. *5*, 296–308.
441. Yao, W. (2019). IDDF2019-ABS-0277 KLF13 regulates CRC development via cholesterol biosynthesis. In Basic Gastroenterology, (BMJ Publishing Group Ltd and British Society of Gastroenterology), pp. A28–A28.
442. Zabidi, M.A., and Stark, A. (2016). Regulatory Enhancer-Core-Promoter Communication via Transcription Factors and Cofactors. Trends Genet. *32*, 801–814.
443. Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat. Genet. *39*, 1512–1516.
444. Zentner, G.E., and Scacheri, P.C. (2012). The chromatin fingerprint of gene enhancer elements. J. Biol. Chem. *287*, 30888–30896.
445. Zhang, C., and Ma, Y. (2012). Ensemble machine learning: methods and applications (Springer Science & Business Media).
446. Zhang, B., Jia, W.-H., Matsuda, K., Kweon, S.-S., Matsuo, K., Xiang, Y.-B., Shin, A., Jee, S.H., Kim, D.-H., Cai, Q., et al. (2014). Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nat. Genet. *46*, 533–542.
447. Zhang, W., Hart, J., McLeod, H.L., and Wang, H.L. (2005). Differential expression of the AP-1 transcription factor family members in human colorectal epithelial and neuroendocrine neoplasms. Am. J. Clin. Pathol. *124*, 11–19.
448. Zhang, W., Bojorquez-Gomez, A., Velez, D.O., Xu, G., Sanchez, K.S., Shen, J.P., Chen, K., Licon, K., Melton, C., Olson, K.M., et al. (2018). A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat. Genet. *50*, 613–620.
449. Zhang, X., Yang, J.-J., Kim, Y.S., Kim, K.-Y., Ahn, W.S., and Yang, S. (2010). An 8-gene signature, including methylated and down-regulated glutathione peroxidase 3, of gastric cancer. Int. J. Oncol. *36*, 405–414.
450. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS ONE *9*, e78644.
451. Zhu, J., Xiong, G., Fu, H., Evers, B.M., Zhou, B.P., and Xu, R. (2015). Chaperone hsp47 drives malignant growth and invasion by modulating an ECM gene network. Cancer Res. *75*, 1580–1591.
452. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. Royal Statistical Soc. B *67*, 301–320.
453. Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. Genome Biol. *19*, 217.
454. Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc Natl Acad Sci USA *111*, 996–1001.